

Integrating Domain Knowledge
for Robust and Interpretable Deep Neural Networks
in Facial Affect Recognition

Ines Sophia Rieger

Integrating Domain Knowledge for Robust and Interpretable Deep Neural Networks in Facial Affect Recognition

Ines Sophia Rieger



Otto-Friedrich-Universität Bamberg

Fakultät Wirtschaftsinformatik und Angewandte Informatik

CUMULATIVE DISSERTATION THESIS

*Submitted in partial fulfillment of the requirements for the degree of Doctor
of Natural Sciences (Dr. rer. nat.) of the University of Bamberg*

*Supervisor and
Reviewer*

Prof. Dr. Ute Schmid
Chair of Cognitive Systems

Second Reviewer

Prof. Dr. Christian Ledig
Chair of Explainable Machine Learning

*Head of Doctoral
Committee*

Prof. Dr. Maximilian E. Schüle

Submitted

12.03.2026

Defended

20.07.2026

Dieses Werk ist als freie Onlineversion über das Forschungsinformationssystem (FIS; <https://fis.uni-bamberg.de>) der Universität Bamberg erreichbar.
Das Werk steht unter der CC-Lizenz CC BY.

Lizenzvertrag: Creative Commons Namensnennung 4.0
<https://creativecommons.org/licenses/by/4.0/>



URN: urn:nbn:de:bvb:473-irb-116940x
DOI: <https://doi.org/10.20378/irb-116940>

Ines Sophia Rieger
*Integrating Domain Knowledge
for Robust and Interpretable Deep Neural Networks
in Facial Affect Recognition*
Cumulative Dissertation Thesis, 2026
Supervisor: Prof. Dr. Ute Schmid

Otto-Friedrich-Universität Bamberg
Fakultät Wirtschaftsinformatik und Angewandte Informatik
Bamberg, Germany

Abstract

Facial affect recognition is a key component of human-centered AI, enabling systems to respond appropriately to human nonverbal signals. The deployment of deep learning for facial affect recognition is critically hindered by vulnerability to data-driven biases and a lack of transparency. This thesis addresses these challenges through a multi-faceted framework, structured along the CRISP-ML(Q) machine learning lifecycle, integrating domain knowledge to build more robust and interpretable systems.

First, it quantifies demographic bias through extensive subject-level metadata annotation, uncovering performance gaps across demographic groups in leading facial expression recognition (FER) classifiers. To mitigate demographic and multi-label class imbalance, an optimization-based data balancing technique, MLOPT, is introduced to improve performance, fairness, and generalization. Second, the training of deep learning models is guided by injecting anatomical knowledge about facial Action Units (AUs), that is, localized facial muscle activations defined by the Facial Action Coding System (FACS). This is achieved through the CorrLoss regularization term and a Facial Movement Constraint Network (FMC-Net), which together improve model robustness and predictive performance. Third, interpretability is enhanced through a fuzzy-logic ensemble that combines expert rules with model predictions. This dual approach enables self-validation of the models by flagging uncertain outputs while providing inherently interpretable behavior through logical rules. Finally, a new evaluation framework, RelEx, is proposed to assess the physiological plausibility of trained models by quantifying the alignment between feature attributions and expert-defined facial regions.

By addressing these challenges with targeted solutions, this research demonstrates that systematically integrating domain knowledge moves FER beyond performance optimization. The resulting framework provides a foundation for developing AI systems that are not only high-performing but also robust and interpretable.

Zusammenfassung

Titel: Integration von Domänenwissen für robuste und interpretierbare tiefe neuronale Netze in der Erkennung von Gesichtsausdrücken

Affekterkennung in Gesichtern ist ein zentraler Baustein menschenzentrierter KI und ermöglicht es Systemen, angemessen auf menschliche nonverbale Signale zu reagieren. Der Einsatz von Deep Learning zur Gesichtsaffekterkennung wird durch die Anfälligkeit für datengetriebene Verzerrungen (Bias) und mangelnde Transparenz erheblich behindert. Diese Dissertation begegnet diesen Herausforderungen, indem sie ein vielschichtiges Framework vorstellt, das sich am CRISP-ML(Q)-Lebenszyklus für maschinelles Lernen orientiert und Domänenwissen integriert, um robustere und interpretierbarere Systeme zu entwickeln.

Erstens quantifiziert die Arbeit demografische Verzerrungen durch umfangreiche Metadaten Annotationen auf Probandenebene und deckt dabei Leistungsunterschiede zwischen demografischen Gruppen in führenden Klassifikatoren zur Erkennung von Gesichtsausdrücken (Facial Expression Recognition, FER) auf. Um sowohl demografische Ungleichgewichte wie auch multilabel Klassenungleichgewichte zu mindern, wird ein optimierungsbasiertes Verfahren (MLOPT) zur Datenbalancierung eingeführt, welches Leistung, Fairness und Generalisierungsfähigkeit verbessern kann. Zweitens wird der Lernprozess von Deep Learning Modellen durch die Einbindung anatomischen Wissens über Facial Action Units (AUs)–lokale Muskelaktivierungen im Gesicht, wie sie im Facial Action Coding System (FACS) definiert sind–gesteuert. Dies wird durch den Regularisierungsterm CorrLoss und ein Facial Movement Constraint Network (FMC-Net) erreicht, die zusammen die Robustheit und Leistung der Modelle erhöhen können. Drittens wird zur Steigerung von Interpretierbarkeit ein Fuzzy-Logik-Ensemble entwickelt, das Expertenregeln mit Modellvorhersagen kombiniert. Dieser duale Ansatz ermöglicht eine Selbstvalidierung der Modelle, indem unsichere Vorhersagen sichtbar werden, und bietet durch die Verwendung logischer Regeln ein inhärent interpretierbares Verhalten. Schließlich wird mit RelEx ein neues Evaluationsframework vorgeschlagen, welches die physiologische Plausibilität von Modellerklärungen bewertet, indem es die Übereinstimmung zwischen Merkmalsrelevanzen und expertenbasiert definierten Gesichtsregionen quantifiziert.

Indem diese unterschiedlichen Herausforderungen mit gezielten Lösungen adressiert werden, zeigt die Arbeit, dass die systematische Integration von Domänenwissen die FER über die reine Leistungsoptimierung hinausführt. Das resultierende Rah-

menwerk bildet eine Grundlage für die Entwicklung von KI-Systemen, die nicht nur leistungsstark, sondern auch robust und interpretierbar sind.

Acknowledgement

This dissertation would not have been possible without the guidance, support, and encouragement of many individuals. I would like to take this opportunity to express my deepest gratitude to all who have accompanied me on this academic journey.

First and foremost, I extend my heartfelt gratitude to my supervisors, Prof. Dr. Ute Schmid and Prof. Dr. Christian Ledig. Your mentorship, unwavering support, and insightful feedback were the cornerstones of this thesis. Ute, your vision and trust gave me the freedom and confidence to explore complex ideas beyond the immediate scope of the project. Christian, your meticulous feedback and critical questions consistently challenged me to refine my work to a higher standard. Thank you both for fostering an environment of intellectual curiosity and for your patience and guidance throughout this entire process.

I extend my sincere gratitude to the Fraunhofer Institute for Integrated Circuits (IIS) for providing a supportive and innovative research environment. This thesis was developed as part of the cross-institutional project Transparent Medical Expert Companion (TraMeExCo), funded by the German Federal Ministry of Education and Research (BMBF) from 2018 to 2021. TraMeExCos mission to make AI systems in healthcare more robust and interpretable provided the direction for the research presented here. Within this collaborative project, my role focused on robust and transparent facial analysis for pain recognition, a topic I expanded to address the broader challenges in facial affect recognition. I would like to thank Bettina Finzel from the University of Bamberg for the great collaboration and insightful discussions throughout the TraMeExCo project.

I am also very grateful to my colleagues at Fraunhofer IIS. Special thanks go to my group leaders throughout this time: Jens-Uwe Garbas, Dominik Seuss, and Nina Holzer. They provided steadfast support and leadership. I owe a particular debt of gratitude to my Fraunhofer colleagues, Jaspar Pahl and Robert Obermeier. Our countless discussions and collaborations were instrumental to my progress, and I learned a great deal from working with you. I would also like to thank the diligent students who contributed to data annotation, preprocessing, and model training. Your work was invaluable. Finally, I extend my gratitude to the various co-authors and collaborators who contributed to the publications that comprise this thesis.

Finally, on a personal note, I could not have completed this journey without the love and unwavering support of my family and friends. Thank you for everything.

List of Publications

The following publications contain the scientific contributions, visualizations and results produced as part of this doctoral research. For clarity, each included publication is assigned a unique identifier P1–P7, which is used consistently throughout the thesis to refer to the corresponding paper. The publications are categorized by venue and listed in reverse chronological order, with the most recent appearing first.

I. The following publications are published in peer-reviewed conferences:

- P1 Ines Rieger** (2024). “Investigating the Regularization of Deep Neural Networks for Affect Recognition With Relevance-Guided Local Explanations”. In: *Artificial Intelligence: Methodology, Systems, and Applications: 19th International Conference, AIMSIA 2024*. Lecture Notes in Computer Science. Springer, pp. 122–127. DOI: [10.1007/978-3-031-81542-3_10](https://doi.org/10.1007/978-3-031-81542-3_10)
- P2 Bettina Finzel, Ines Rieger, Simon Kuhn, and Ute Schmid** (2023). “Domain-Specific Evaluation of Visual Explanations for Application-Grounded Facial Expression Recognition”. In: *Machine Learning and Knowledge Extraction. CD-MAKE 2023*. Lecture Notes in Computer Science. Springer, pp. 31–44. DOI: [10.1007/978-3-031-40837-3_3](https://doi.org/10.1007/978-3-031-40837-3_3)
- P3 Jaspar Pahl, Ines Rieger, Anna Möller, Thomas Wittenberg, and Ute Schmid** (2022). “Female, white, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces”. In: *Proc. of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, 973–987. DOI: [10.1145/3531146.3533159](https://doi.org/10.1145/3531146.3533159)¹
- P4 Ines Rieger, Jaspar Pahl, Bettina Finzel, and Ute Schmid** (2022). “CorrLoss: Integrating Co-Occurrence Domain Knowledge for Affect Recognition”. In: *Proc. of the 26th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 798–804. DOI: [10.1109/ICPR56361.2022.9956319](https://doi.org/10.1109/ICPR56361.2022.9956319)
- P5 Ines Rieger, René Kollmann, Bettina Finzel, Dominik Seuss, and Ute Schmid** (2020a). “Verifying Deep Learning-based Decisions for Facial Expression Recognition”. In: *Proc. of the 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pp. 139–144. URL: <https://www.esann.org/sites/default/files/proceedings/2020/ES2020-49.pdf>

¹Both authors Pahl and Rieger contributed equally to this research.

II. The following publication is published in a peer-reviewed conference workshop:

P6 Ines Rieger, Jaspar Pahl, and Dominik Seuss (2020b). “Unique Class Group Based Multi-Label Balancing Optimizer for Action Unit Detection”. In: *Proc. of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG), Workshop on Affect Recognition in-the-wild: Uni/Multi-Modal Analysis and VA-AU-Expression Challenges*. IEEE, pp. 619–623. DOI: [10.1109/FG47880.2020.00101](https://doi.org/10.1109/FG47880.2020.00101)²

III. The following publication is published in a peer-reviewed journal:

P7 Ines Rieger, Jaspar Pahl, and Ute Schmid (2025). “FMC-Net: A Human-Guided Deep Learning Framework for Adaptable and Transparent Facial Expression Recognition in Real-World Scenarios”. In: *Applied Intelligence* 55.18. DOI: [10.1007/s10489-025-07017-9](https://doi.org/10.1007/s10489-025-07017-9)

A detailed account of my scientific and written contributions is provided in Appendix [B](#).

²This approach won third place in the Affective Behavior Analysis in-the-wild (ABAW) challenge (Kollias et al., [2020](#)).

Contents

I	Synopsis of Dissertation	1
1	Introduction	3
1.1	Research Challenges	3
1.1.1	Model Robustness	3
1.1.2	Model Interpretability	4
1.2	Research Approach: Integrating Domain Knowledge	5
1.3	Research Questions	5
1.3.1	Phase 1: Data Understanding - Measuring Bias	6
1.3.2	Phase 2: Data Preparation - Mitigating Bias	6
1.3.3	Phase 3: Machine Learning Modeling - Knowledge-Informed Training and Inference	7
1.3.4	Phase 4: Model Evaluation - Validating Model Reasoning	8
1.4	Scientific Contributions	8
1.5	Methodological Considerations	9
1.5.1	Methodological Consistency	10
1.5.2	Reproducibility and Constraints on Open-Sourcing	10
1.6	Structure of Synopsis	11
2	Background and Related Work	13
2.1	Facial Affects: From Action Units to Expressions	13
2.1.1	Facial Action Units	13
2.1.2	Facial Expressions: Basic Emotions and Pain	16
2.2	Domain Knowledge	16
2.2.1	Localization of Facial Action Units	18
2.2.2	Co-Occurrence of Facial Action Units	18
2.2.3	Hierarchical Relationship Between Facial Expressions and Fa- cial Action Units	19
2.2.4	Intensity Scale and Temporal Modality of Facial Expressions	20
2.2.5	Subject Demographic Metadata	21
2.3	Deep Neural Networks	21
2.3.1	Learning Setup	22

2.3.2	Network Architectures	22
2.4	Evaluation Framework	22
2.4.1	Evaluating Model Robustness	23
2.4.2	Evaluating Model Interpretability	25
2.4.3	Correlation Coefficients for Model Regularization	26
3	Datasets and Data Preprocessing	29
3.1	Dataset Acquisition and Management	29
3.2	Dataset Characteristics	29
3.3	Data Preprocessing	34
4	Bias Analysis in Datasets and Models [P3]	37
4.1	Motivation and Related Work	37
4.2	Demographic Metadata Annotation	38
4.3	Summary of Key Results	39
4.3.1	Bias Analysis in Datasets [P3]	39
4.3.2	Bias Analysis in Models [P3]	40
4.4	Summary and Discussion	41
5	Multi-Label Data Balancing: MLOPT and MLOPT-Ext [P6]	43
5.1	Motivation and Related Work	43
5.2	General Methodology	46
5.2.1	MLOPT: Balancing Classes	47
5.2.2	MLOPT-Ext: Balancing Multiple Objectives	51
5.3	Summary of Key Results	52
5.3.1	MLOPT: Superior to Competitive Algorithms on Benchmark Datasets	52
5.3.2	MLOPT: Improved DNN Performance for In-the-wild AU De- tection [P6]	53
5.3.3	MLOPT-Ext: Simultaneous Class and Demographic Balancing	53
5.4	Summary and Discussion	54
6	Knowledge-Informed Model Training: CorrLoss and FMC-Net [P4, P7]	55
6.1	Motivation and Related Work	55
6.2	General Methodology	56
6.2.1	Correlation Constraints	57
6.2.2	Unified Objective and Single-Task Case	58
6.3	Summary of Key Results	58
6.3.1	CorrLoss: AU–AU Constraint Improves Performance and Gen- eralizability in AU Detection [P4]	58

6.3.2	FMC-Net: $AU \leftrightarrow FE$ Constraint Improves Performance for Joint AU-FE Prediction [P7]	59
6.4	Summary and Discussion	60
7	Rule- and Model-Based Ensemble for Transparent Inference: FMC-Net [P7]	63
7.1	Motivation and Related Work	63
7.2	General Methodology	64
7.2.1	Fuzzy $AU \rightarrow FE$ Mapping (Rule-Based Inference)	65
7.2.2	Rule-Model Ensemble and Disagreement-Based Risk	65
7.3	Summary of Key Results: Increasing Robustness and Interpretability for Emotion Recognition [P7]	66
7.4	Summary and Discussion	67
8	Model Evaluation: Relevance-Guided Local Explanations (RelEx) [P1, P2, P5]	69
8.1	Motivation and Related Work	70
8.2	General Methodology	71
8.3	Summary of Key Results	72
8.3.1	Unraveling Decisions of Trained CNNs for Action Unit Detec- tion [P2, P5]	72
8.3.2	Effect of Co-occurrence Regularization on Trained Features for Action Unit Detection [P1]	73
8.4	Summary and Discussion	74
9	Conclusion and Outlook	77
9.1	Summary of Key Findings	77
9.2	Discussion and Outlook	79
	Bibliography	81
II	Publications of Dissertation	93
A	Additional Results	95
A.1	Overview of Tabular Benchmark Datasets	95
A.2	Comparison of MLOPT with Different Objective Functions	96
A.3	Comparison of MLOPT with Competitive Approaches	97
A.4	Evaluation of MLOPT-Ext on BP4D	99
B	Publications	103

B.1	[P1] Investigating the Impact of Neural Network Regularization for Affect Recognition through Quantified Local Explanations	103
B.2	[P2] Domain-Specific Evaluation of Visual Explanations for Application-Grounded Facial Expression Recognition	104
B.3	[P3] Female, White, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces	105
B.4	[P4] CorrLoss: Integrating Co-occurrence Domain Knowledge for Affect Recognition	106
B.5	[P5] Verifying Deep Learning-Based Decisions for Facial Expression Recognition	107
B.6	[P6] Unique Class Group Based Multi-Label Balancing Optimizer for Action Unit Detection	108
B.7	[P7] FMC-Net: A Human-Guided Deep Learning Framework for Adaptable and Transparent Facial Expression Recognition in Real-World Scenarios	109

List of Figures

1.1	Thesis overview	11
4.1	Age distribution of all combined datasets, with hue for dataset. From [P3] (Pahl et al., 2022).	40
4.2	Age distribution of all combined datasets, hue ethnicity (AA: African-American, AS: Asian, EA: Euro-American, HI: Hispanic, O: Other). From [P3] (Pahl et al., 2022).	40
5.1	Distribution of AUs in a merged database comprising Actor Study, Aff-Wild2, CK+, and EmotioNet (Manual).	44
6.1	CorrLoss (d) brings predicted AU–AU correlations closer to target correlations (b) (adapted from [P4] (Rieger et al., 2022)). This is calculated for the FE happy on the BP4D dataset.	62
6.2	Constraint-guided training in FMC-Net for emotion recognition: a multi-task CNN predicts AUs and FEs while a differentiable term encodes AU–FE relationships to regularize shared representations. Adapted from [P7] (Rieger et al., 2025).	62
7.1	FMC-Net inference with a rule–model ensemble. AU probabilities from the AU head are mapped to emotions by a fuzzy, user-editable AU-to-FE rule layer, while a parallel CNN head provides model-based emotion predictions. The ensemble combines both sources to produce the final decision and an interpretable risk score based on their disagreement. Adapted from [P7] (Rieger et al., 2025).	64
8.1	RelEx pipeline. (1) Class-specific saliency map (here: AU4 Brow Lowerer) from any local explainer. (2) Facial landmarks define facial sub-regions. (3) Domain knowledge (AU-localisation) defines AU-specific regions of interest (ROIs). (4) Quantification: the μ -score measures the fraction of relevance inside the ROI. Adapted from Rieger (2024).	71

9.1	Overview of thesis contributions mapped to CRISP-ML(Q). Each research question (RQ) is shown with its method, associated publication(s) (P), and the domain knowledge (DK) (created $\odot$, injected $\ominus$). Arrows indicate improvements per RQ: Performance (P), Interpretability (I), Robustness (R), Generalizability (G), Fairness (F).	77
-----	---	----

List of Tables

2.1	Overview of affective label types used across the included publications.	13
2.2	Overview of the 20 AUs used in this thesis, as defined in FACS (Ekman and Friesen, 1978; Ekman et al., 2002).	15
2.3	Overview of domain knowledge types and their application in the methodological contributions of this thesis.	17
2.4	Logical rules for mapping AUs to basic emotions (Friesen, Ekman, et al., 1983) and pain (Kunz and Lautenbacher, 2014). Adapted from [P7] (Rieger et al., 2025).	19
2.5	Overview of learning setups, target tasks, and network architectures used across the included publications. Abbrev.: Train = model training performed, ML = multi-label; MC = multi-class; MT = multi-task.	21
2.6	Overview of evaluation frameworks used across the included publications. Abbrev.: W = within-dataset; X = cross-dataset; SD-CV = subject-dependent cross-validation; RelEx = quantified saliency maps.	23
2.7	Quantitative evaluation overview across the included publications: evaluated entity, quantitative aspect (one aspect per row), and corresponding metric(s). Abbrev.: PP = predictive performance; RelEx = quantified saliency maps.	24
3.1	Overview of datasets used in this thesis and corresponding publications in which they were employed. Adapted from Pahl et al. (2022), Table 1.	30
4.1	Overview of datasets and available demographic metadata for gender, ethnicity, and age. Adapted from [P3] (Pahl et al., 2022).	38
5.1	Overview of competitive multi-label oversampling (OS) balancing algorithms.	46
A.1	Overview of datasets. MeanIR is the average of the imbalance ratios across all labels.	96
A.2	Comparison of L1, L2, and Kullback-Leibler Divergence (KLDIV). Best values per dataset are in bold.	97

A.3	Comparison of competitive algorithms for all datasets with bounds = 5. Metric is MeanIR.	98
A.4	Comparison of competing algorithms for all datasets with bounds = 10. Metric is MeanIR.	98
A.5	Comparison of competing algorithms for all datasets with % increase = 10. Metric is MeanIR.	99
A.6	Comparison of competing algorithms for all datasets with % increase = 20. Metric is MeanIR.	99
A.7	MLOPT-Ext applied on BP4D dataset. Original (unoptimized) standard deviations are: AU: 0.28155, Ethnicity: 0.3322, Age: 0.1518, Gender: 0.1188. Unoptimized MeanIR: 2.715. Values worse than these original values are <u>underlined</u> ; for each row, the better value between L1 and L2 for each attribute is bolded	101

List of Abbreviations

ABAW	Affective Behavior Analysis in-the-Wild
AI	Artificial Intelligence
AU	Action Units
AUROC	Area Under the Receiver Operating Characteristic Curve
BP4D	BP4D-Spontaneous Database
CCC	Concordance Correlation Coefficient
CNN	Convolutional Neural Network
CRISP-ML(Q)	Cross-Industry Standard Process for Machine Learning with Quality Assurance
DNN	Deep Neural Network
EMFACS	Emotion Facial Action Coding System
FACS	Facial Action Coding System
FE	Facial Expression
FER	Facial Expression Recognition
FMC-Net	Facial Movement Constraint Network
Grad-CAM	Gradient-weighted Class Activation Mapping
LRP	Layer-wise Relevance Propagation
MeanIR	Mean Imbalance Ratio
ML	Machine Learning
MLSMOTE	Multi-Label Synthetic Minority Over-sampling Technique
MLOPT	Multi-Label Balancing Optimizer
MLOPT-Ext	Extended Multi-Label Balancing Optimizer

PCC Pearson Correlation Coefficient

RelEx Relevance-Guided Local Explanations

ROI Region of Interest

XAI Explainable Artificial Intelligence

Part I

Synopsis of Dissertation

Introduction

Deep learning has become a dominant force in computer vision (LeCun et al., 2015; Zhang et al., 2023a), particularly in the domain of facial affect recognition, the task of identifying human emotional states from facial images (Picard, 2000; Leong et al., 2023). This thesis addresses the development of robust and interpretable models for two fundamental affective tasks: Facial Expression Recognition (FER), which focuses on visible expressions of emotions or feelings through facial movements, and Action Unit (AU) detection. AUs, defined by the Facial Action Coding System (FACS) (Ekman and Friesen, 1978), are the modular, anatomical movements of facial muscles (e.g., raising an eyebrow) that compose complex facial expressions (FEs). Together, AU detection and FER provide complementary perspectives on facial affect: FER captures high-level categorical expressions, whereas AUs offer an anatomy-grounded, decomposable representation. This combination enables domain-knowledge integration and interpretability by linking predictions to plausible facial muscle activations and by leveraging AU–FE task relationships. While deep neural networks (DNNs) have achieved remarkable performance on FER tasks (Li et al., 2023; Yu et al., 2023; Yang et al., 2021; Dehshibi and Masip, 2024), their application in real-world, high-stakes scenarios is hindered by fundamental challenges related to model robustness and interpretability (Leong et al., 2023; Nourivandi et al., 2024; Li and Deng, 2020).

1.1 Research Challenges

Despite their strong performance, the deployment of deep learning models for FER faces two fundamental research challenges: ensuring robustness against variations in real-world data and providing interpretability to demystify their black box nature. These challenges are not merely technical; they are essential to developing trustworthy AI, a concern emphasized by emerging regulatory frameworks such as the EU AI Act (Panigutti et al., 2023).

1.1.1 Model Robustness

A primary challenge is robustness, that is, a models ability to generalize and maintain performance when faced with real-world data variations that were not present or

fully represented in the training data (Li and Deng, 2020). This is particularly acute in FER: models often fail to generalize across different demographics (e.g., age, gender, ethnicity) or under challenging imaging conditions such as occlusions, head pose, or varying illumination ([P3] (Pahl et al., 2022); (Nourivandi et al., 2024)). These failures are frequently rooted in selection biases in datasets, leading to poor fairness, i.e., inequitable performance across subgroups (Chapter 4). This thesis addresses these issues by developing data-centric methods to mitigate data imbalance (Chapter 5) and model-centric methods to enforce domain-specific constraints (Chapter 6).

Robustness in this thesis: Robustness refers to a models ability to generalize by maintaining high performance on unseen images from known classes, even when those images exhibit variations in style, acquisition conditions, or subject characteristics. In addition, robustness is considered with respect to demographic subpopulations: a robust model should avoid large performance drops for specific groups. In this sense, robustness and fairness are closely linked, and fairness is operationalized as reduced performance disparities across demographic subgroups.

1.1.2 Model Interpretability

The second fundamental challenge is the opaque, black-box nature of DNNs (Guidotti et al., 2018; Schmid, 2024). In high-stakes domains like healthcare, this lack of transparency inhibits user trust and prevents meaningful human oversight. To address this, the field of explainable AI (XAI) provides methods to illuminate a models decision-making process (Watson, 2022; Van Lent et al., 2004). This thesis contributes novel XAI techniques that are not domain-agnostic, but are instead grounded in specific domain knowledge. The goal is to move beyond generic explanations by enabling domain-specific ensembles (Chapter 7) and quantifying the alignment of model reasoning with anatomical expertise (Chapter 8).

Interpretability in this thesis: Interpretability refers to the degree to which a models predictions and internal reasoning can be understood by a human in terms of meaningful domain concepts. Consequently, interpretability is sought by making the decision process more transparent and by providing explanations that go beyond performance metrics, linking model behavior to facial anatomy (AUs, regions) and psychologically motivated rules.

1.2 Research Approach: Integrating Domain Knowledge

To address the challenges of robustness and interpretability, this dissertation moves beyond purely data-driven methods toward human-guided systems in which domain-specific insights constrain and shape the model lifecycle (Dash et al., 2022; Bayer et al., 2022; Shneiderman, 2022). In this thesis, the systematic integration of such knowledge is structured according to the CRISP-ML(Q) framework (Studer et al., 2021), an industry-standard process model for developing machine learning applications from data understanding to final evaluation.

In FER, relevant domain knowledge stems from psychological and physiological research (see Section 2.2). Frameworks such as FACS provide a structured account of anatomical muscle movements and their co-occurrence patterns. By explicitly embedding this information—from AU localization and co-occurrences to hierarchical AU–emotion relationships and demographic distributions—the goal is to create models that are not only more accurate but also better aligned with established human expertise.

Domain knowledge integration in this thesis: Domain knowledge integration denotes the structured incorporation of field-specific knowledge (e.g., anatomical constraints, statistical regularities, and expert heuristics) into the machine learning lifecycle. This includes using such knowledge to guide data design and curation, the choice of learning objectives and regularization terms, training procedures, and evaluation protocols.

1.3 Research Questions

Building on the research approach of integrating domain knowledge, this thesis is driven by the central research question:

How can domain knowledge be integrated into deep neural networks to improve robustness and interpretability in facial expression recognition?

Anchored in CRISP-ML(Q), the thesis targets the phases of data understanding, model development, and evaluation (deliberately excluding deployment and monitoring) and addresses five sub-questions mapped to this scope.

1.3.1 Phase 1: Data Understanding - Measuring Bias

A prerequisite for building robust and fair models is understanding the biases within the data. While class imbalances are readily analyzed, the lack of demographic metadata in FER datasets makes it difficult to assess fairness across groups (Zhang et al., 2014; Girard et al., 2017; Nourivandi et al., 2024). This motivates the first research question:

***RQ1:** How can fairness across demographic groups be measured in facial expression recognition datasets and trained models?*

This question is addressed through the systematic annotation of demographic attributes (e.g., age, gender, ethnicity) and a subsequent bias analysis across benchmark datasets and models, as detailed in Chapter 4 ([P3] Pahl et al. (2022)). The demographic annotation itself constitutes domain knowledge that enables this analysis.

1.3.2 Phase 2: Data Preparation - Mitigating Bias

Identifying dataset and model bias is insufficient; it must be mitigated. A key challenge is that FER datasets often have multi-label class imbalances (multiple AUs co-occurring) and demographic imbalances. Class imbalance is amplified from the natural variability in the frequency of FEs or AUs (Karkkainen and Joo, 2021). Demographic imbalances in FER datasets often stem from their limited number of subjects, as collecting and annotating high-quality affective data is resource-intensive (Seuss et al., 2019). Existing balancing methods typically fail to address these multiple, often conflicting, objectives in a way that is suitable for deep learning (see Section 5.1). This motivates the second research question:

***RQ2:** How can multi-label datasets be balanced across multiple objectives in a way that remains suitable for training deep neural networks?*

This question is addressed by the multi-label balancing optimizer (MLOPT) and its extension MLOPT-Ext for multiple objectives (e.g., class frequency and demographic representation). MLOPT-Ext is a non-synthetic, optimization-based balancing framework. This data-level mitigation is model-agnostic and can improve the training data before any architectural or loss-level changes. The approach is described in Chapter 5 ([P6] Rieger et al. (2020b)).

1.3.3 Phase 3: Machine Learning Modeling - Knowledge-Informed Training and Inference

Beyond data preparation, the model itself can be improved by integrating domain knowledge. The third and fourth research questions concern (1) constraints during training and (2) structured reasoning at inference-time.

(1) Training Stage: To improve models beyond data pre-processing, this thesis investigates model-centric constraints during training. The co-occurrence of AUs and interrelationships between different AUs and emotion represents powerful domain knowledge. While prior methods have modeled these relationships using indirect approaches like graph networks or hypothesis spaces (see Section 6.1), this thesis argues that a direct constraint is an effective mechanism for improving model generalization, motivating the third research question:

RQ3: *How can relationships between classes and tasks be enforced during training?*

This question is addressed by the CorrLoss method and the facial movement constraint network (FMC-Net), which introduce correlation-based regularization terms that encode inter-class and inter-task relationships directly in the loss function. These correlations use domain knowledge as statistically stable structure that is expected to persist beyond a single dataset. These methods are presented in Chapter 6 ([P4] Rieger et al. (2022), [P7] Rieger et al. (2025)).

(2) Inference Stage: While RQ3 injects knowledge into training, RQ4 integrates domain knowledge at inference-time to increase transparency and provide actionable risk signals. A promising strategy for improving robustness is the use of ensemble models (see Section 7.1). However, existing ensembles in FER are typically data-driven, combining multiple neural networks. This overlooks the potential of integrating structured, expert-driven logic directly into the inference process. This thesis investigates a hybrid approach, motivating the fourth research question:

RQ4: *How can an expert-informed ensemble be designed for facial expression recognition?*

This question is addressed by a fuzzy logic-based ensemble that fuses probabilistic model outputs with rule-based emotion inference derived from AUs, enabling transparent reasoning and a disagreement-based risk assessment for emotion recognition. This method is described in Chapter 7 ([P7] Rieger et al. (2025)).

1.3.4 Phase 4: Model Evaluation - Validating Model Reasoning

Finally, a trustworthy model must be right for the right reasons (Schmid, 2023; Watson, 2022). It is not enough for a model to be accurate; its internal reasoning must be understood, especially for sensitive applications (see Section 8.1). This requires moving beyond standard performance metrics to quantitatively evaluate a model's explanations with XAI techniques. This motivates the fifth research question:

***RQ5:** How can the alignment between model predictions and physiologically plausible Action Unit locations be quantified for model validation?*

This question is addressed by the relevance-guided local explanations (RelEx) framework. This framework quantitatively compares saliency maps against known AU localization derived from facial physiology ([P2] Finzel et al. (2023), [P5] Rieger et al. (2020a)). RelEx is also applied to the CorrLoss regularization to analyze the training effects ([P1] Rieger (2024)). The method is described in Chapter 8.

Overall, the thesis follows an evaluation-driven loop: demographic bias is first identified, then mitigated via balancing, physiologically/statistically grounded structure is enforced during training, and model attention is finally validated with respect to plausible facial regions.

1.4 Scientific Contributions

This thesis answers the research questions through the following key scientific contributions:

Bias Analysis (RQ1):

- This work presents a systematic analysis of demographic bias in AU and FER benchmark datasets. This was enabled by annotating previously missing meta-data (age, gender, ethnicity, occlusion), providing a foundational contribution for evaluating and improving model fairness in the field ([P3] Pahl et al. (2022)).

Mitigating Dataset Imbalances (RQ2):

- To mitigate the biases identified in RQ1, this work introduces the MLOPT framework, a non-synthetic, optimization-based approach for data balancing ([P6] Rieger et al. (2020b)). Initially designed for multi-label data, MLOPT was extended to MLOPT-Ext, the first algorithm to simultaneously balance

both class labels and demographic attributes, directly incorporating domain knowledge to train fairer and more robust models. The core MLOPT approach was validated by winning third place in the international the Affective Behavior Analysis in-the-wild (ABAW) challenge at the IEEE Conference on Face and Gesture Recognition 2020 Kollias et al. (2020).

Knowledge-Informed Model Training (RQ3):

- This thesis introduces CorrLoss, a correlation-based regularization method that directly integrates class co-occurrence patterns into the DNN loss function, improving generalization in multi-label FER ([P4] Rieger et al. (2022)). Building on this idea, the multi-task FMC-Net extends CorrLoss to jointly model hierarchical target relationships between AUs and emotions, yielding an adaptable training approach that improves performance while reducing task interference ([P7] Rieger et al. (2025)).

Rule-Based Ensemble with Fuzzy Logic (RQ4):

- This work develops an expert ensemble approach for emotion recognition that combines model-based predictions with rule-based predictions derived from fuzzy logic ([P7] Rieger et al. (2025)). By encoding established psychological knowledge in interpretable rules and contrasting them with model outputs, the ensemble enhances interpretability, is easily adapted to new domains via its rule base, and provides a disagreement-based risk signal for identifying unreliable predictions.

Quantifying Saliency Maps for Model Evaluation (RQ5):

- To validate model reasoning, this work introduces RelEx, a relevance-guided framework that quantitatively evaluates AU saliency maps by measuring their alignment with anatomical domain knowledge ([P2] Finzel et al. (2023), [P5] Rieger et al. (2020a)). RelEx complements and goes beyond conventional performance metrics by assessing whether models base their decisions on physiologically plausible facial regions, and is used to provide the first quantitative analysis of how co-occurrence regularization (CorrLoss) shapes a model's learned features ([P1] Rieger (2024)).

A detailed account of the scientific and written contributions for each publication is provided in Appendix B.

1.5 Methodological Considerations

Beyond the individual scientific contributions, this cumulative thesis also follows shared methodological principles to ensure coherence, comparability, and trustwor-

thiness across all included publications. The following subsections summarize key aspects of methodological consistency and discuss constraints and choices related to reproducibility and open-sourcing.

1.5.1 Methodological Consistency

To ensure coherence and enable systematic comparison across the cumulative contributions of this thesis, consistency was maintained in the reuse of benchmark datasets, DNN architectures, incorporated domain knowledge, and evaluation metrics across multiple publications.

Overviews relevant to this consistency are provided in several tables: Table 2.1 outlines the affective classes (tasks) referenced or used across the publications, Table 2.3 provides an overview of the domain knowledge types and their application in the methodological contributions, Table 2.5 summarizes the learning setups and network architectures used, Table 2.6 presents the various evaluation frameworks applied, Table 2.7 offers a comprehensive summary of the performance metrics used in each publication, and Table 3.1 summarizes the datasets used per publication.

While task-specific requirements and venue constraints led to some variation in the reported metrics, key indicators such as the F1-score are used throughout as primary measures of predictive performance. Where possible, newly proposed methods (e.g., CorrLoss, MLOPT-Ext) were evaluated on shared datasets with overlapping metric sets to support comparability. These steps ensure methodological traceability across different experimental contexts.

1.5.2 Reproducibility and Constraints on Open-Sourcing

Transparency and reproducibility are essential pillars of trustworthy research, especially in the context of trustworthy AI. Wherever possible, this thesis therefore builds on publicly available benchmark datasets and standard deep learning architectures, and reports evaluation protocols and hyperparameters in detail.

This work was conducted in collaboration with the Fraunhofer Institute for Integrated Circuits IIS under a formal agreement that imposes strict confidentiality and intellectual property policies. These policies currently restrict the public release of the full source code and of the annotated demographic metadata for internally processed datasets. To mitigate this limitation, the methods, loss formulations, and evaluation procedures are documented in detail in the associated publications and in this synopsis, so that the experiments can be reproduced in principle using the described configurations and publicly available datasets. Where feasible and compatible with contractual obligations, I am happy to share additional technical details

and to support interested researchers in reproducing or extending the presented work.

1.6 Structure of Synopsis

The synopsis is structured as follows: Chapter 2 provides background and related work on AUs, FEs, domain-specific knowledge, evaluation settings, applied metrics, DNN architectures, and classifiers using DNNs. Chapter 3 presents the datasets and preprocessing techniques.

This is followed by the contribution chapters, an overview is visualized in Figure 1.1.

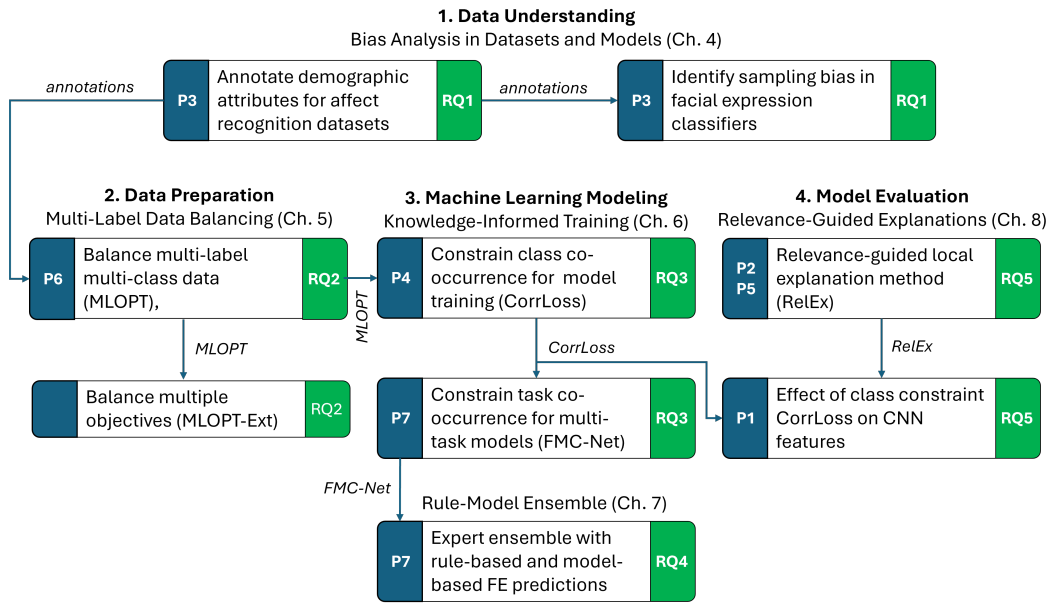


Fig. 1.1.: Thesis overview along CRISP-ML(Q) process steps. The figure summarizes how the research questions (RQ) are connected to the scientific publications (P) and in which chapters the contributions are presented and discussed. Labeled unidirectional arrows indicate how the contributions build on each other.

Each of the contribution chapters follows a structured format: a motivation and related work section, a section describing the proposed method, followed by a section of the summarized results and a discussion section. Chapter 4 presents the annotation of demographic attributes and a subsequent bias analysis in models and datasets. Chapter 5 presents the developed balancing algorithm for dealing with class and demographic imbalances, Chapter 6 presents methods for correlation-informed model training, Chapter 7 describes the ensemble approach, and Chapter 8 introduces the model evaluation framework. Finally, the conclusion and outlook in Chapter 9 discusses the implications of the research and potential future directions.

Appendix [B](#) provides a detailed account of my scientific and written contributions, together with the full texts.

Background and Related Work

2.1 Facial Affects: From Action Units to Expressions

This section provides an overview of facial affects, starting with the low-level anatomical components defined as Action Units (AUs) and extending to higher-level facial expressions such as basic emotions and pain. These affective classes serve as the primary labels used for training, evaluating, and interpreting models throughout this thesis.

Table 2.1 summarizes which types of affective classes were used in each of the included publications.

Tab. 2.1.: Overview of affective label types used across the included publications.

Paper	Main method(s)	Affective label type(s)
P1	Regularized model training (CorrLoss); saliency map quantification (RelEx)	Emotion-related AUs
P2	Saliency map quantification (RelEx)	Pain-related AUs
P3	Bias analysis of benchmark datasets and models	Emotion-related AUs; discrete emotions
P4	Regularized model training (CorrLoss)	Emotion-related AUs
P5	Saliency map quantification (RelEx)	Pain-related AUs
P6	Multi-label class balancing (MLOPT)	AUs
P7	Regularized multi-task model training (FMC-Net); expert ensemble with fuzzy logic	Emotion-related AUs; discrete emotions

2.1.1 Facial Action Units

The anatomical constraints of the facial muscles and tissues dictate the range and variability of possible facial movements. Key muscle groups, such as the zygomaticus major (responsible for lifting the corners of the mouth during smiling), orbicularis oculi (involved in movements around the eyes), and frontalis (raising the eyebrows and wrinkling the forehead), along with the elasticity and texture of the skin, shape distinct facial expressions. These anatomical insights were derived from extensive

psychological studies of human facial expressions, and formalized as Action Units (AUs) in the *Facial Action Coding System (FACS)* by Ekman and Friesen (1978), which was later substantially updated by Ekman et al. (2002).

A total of 44 AUs have been defined in FACS, 30 of which are directly related to one or more facial muscles, while the remaining AUs capture head and eye movements. Of the anatomically related AUs, 12 correspond to upper facial movements, and 18 correspond to lower facial movements. Accurate annotation of these AUs requires the expertise of certified FACS coders due to the complexity and subtlety of facial muscle activations and their combinations. This process is time consuming and costly, resulting in a limited number of annotated datasets with only a subset of AUs annotated, which may vary from dataset to dataset (see Section 3).

Table 2.2 provides an overview of the AU classes used for AU detection models in this thesis, including AU codes, names, and the associated facial muscles. The AUs were selected based on their relevance to basic emotions and pain-related expressions. Note that the AU codes in FACS are not numbered consecutively due to their evolution over time. In addition, the AU codes are grouped by facial region, such as upper or lower facial muscles, and there is no strict numerical order between the groupings. Throughout the thesis, AU codes are used as identifiers for the sake of conciseness.

Tab. 2.2.: Overview of the 20 AUs used in this thesis, as defined in FACS (Ekman and Friesen, 1978; Ekman et al., 2002).

AU Code	AU Name	Facial muscle(s) and description
01	Inner Brow Raiser	Frontalis (pars medialis): raises the inner part of the eyebrows.
02	Outer Brow Raiser	Frontalis (pars lateralis): raises the outer part of the eyebrows.
04	Brow Lowerer	Depressor supercilii, corrugator supercilii: pulls the eyebrows down and inward.
05	Upper Lid Raiser	Levator palpebrae superioris: lifts the upper eyelid.
06	Cheek Raiser	Orbicularis oculi (pars orbitalis): lifts the cheeks and creates crow's feet around the eyes.
07	Lid Tightener	Orbicularis oculi (pars palpebralis): tightens the skin around the eyes.
09	Nose Wrinkler	Levator labii superioris alaeque nasi: wrinkles the nose.
10	Upper Lip Raiser	Levator labii superioris: raises the upper lip.
11	Nasolabial Deepener	Zygomaticus minor: pulls the skin around the nose and upper lip, deepening the nasolabial fold.
12	Lip Corner Puller	Zygomaticus major: pulls the corners of the mouth upwards, as in smiling.
14	Dimpler	Buccinator: pulls the lip corners laterally, sometimes producing dimples in the cheeks.
15	Lip Corner Depressor	Depressor anguli oris (triangularis): pulls the corners of the mouth downward.
16	Lower Lip Depressor	Depressor labii inferioris: pulls the lower lip downward.
17	Chin Raiser	Mentalis: pushes the chin upwards and wrinkles the chin.
20	Lip Stretcher	Risorius with platysma: stretches the lips horizontally.
23	Lip Tightener	Orbicularis oris: tightens the lips.
24	Lip Pressor	Orbicularis oris: presses the lips together.
25	Lips Part	Often due to activity of depressor labii inferioris and/or relaxation of mentalis and orbicularis oris: separates the lips.
26	Jaw Drop	Relaxation of masseter, temporalis and internal pterygoid: lowers the jaw.
27	Mouth Stretch	Pterygoids and digastric: stretches the mouth open.

2.1.2 Facial Expressions: Basic Emotions and Pain

Facial expressions (FEs) refer to the visible expressions of emotions or feelings through facial movements and are a fundamental part of nonverbal communication and a primary method for conveying social information between humans (Darwin, 1872). Facial affects refers to the underlying emotional or affective states that are expressed through FEs. This thesis concentrates on the recognition of visible expressions.

Various annotation systems exist to decode and classify FEs, capturing both discrete and continuous aspects of emotion. In particular, this thesis emphasizes the recognition of basic emotions and pain, which are universally understood and widely used in research and clinical contexts (Leong et al., 2023; Ekman and Friesen, 1971; Kunz and Lautenbacher, 2014).

The recognition of *basic emotions*—anger, disgust, fear, happiness, sadness and surprise—is central to facial expression recognition (FER). These emotions, first introduced by Paul Ekman, are considered biologically innate and recognizable across cultures (Ekman, 1992; Ekman and Friesen, 1971). The term *basic* not only highlights their biological foundation, but also refers to their discrete-emotional perspective (Ekman and Cordaro, 2011). In contrast, more complex emotional states can be captured by continuous annotation systems. Such systems include compound emotions like *angrily disgusted* (Li et al., 2023) and dimensional models that consider variables such as valence (positive vs. negative emotions), arousal (intensity), and novelty (Posner et al., 2005). Despite these more nuanced approaches, basic emotions remain the most prevalent annotation in FER due to their clarity and ease of implementation in computational systems (Leong et al., 2023).

In addition to basic emotions, expressions of pain are another focus of this thesis, especially in clinical contexts where facial cues provide critical nonverbal indicators of a patient’s discomfort. Specific AUs, such as eyebrow lowering and eye tightening, are closely associated with pain and provide a reliable method for nonverbal pain assessment (Craig et al., 1992; Kunz and Lautenbacher, 2014).

2.2 Domain Knowledge

This section provides an overview of the domain knowledge related to AUs, FEs, and pain, which are incorporated into methodological contributions described in the Chapters 4, 5, 6, 7, and 8.

Table 2.3 summarizes how the different types of domain knowledge introduced in Section 2.2 are integrated into the methods and experiments of the included publications.

Tab. 2.3.: Overview of domain knowledge types and their application in the methodological contributions of this thesis.

Domain knowledge type	Application in methods and publications
AU localization (Section 2.2.1)	Region-level AU knowledge used in RelEx to quantify whether models focus on physiologically plausible facial regions in their saliency maps [P1, P2, P5].
AU co-occurrence (Section 2.2.2)	Co-occurrence statistics between AUs used as target correlations in correlation-based regularization (CorrLoss) for model training [P1, P4, P7].
Hierarchical AU–FE relationships (Section 2.2.3)	Structured relationships between AUs and discrete emotions used to design and regularize the multi-task FMC-Net and to define the fuzzy rule base for the expert ensemble [P7].
Intensity scale and temporal modality (Section 2.2.4)	Temporal modality used in model evaluation [P3] and the intensity scale to parameterize the fuzzy-logic integration in the expert ensemble [P7].
Demographic metadata (Section 2.2.5)	Demographic annotations (e.g., age, gender, ethnicity, occlusion) used for bias analysis of datasets and models [P3] and as attributes in multi-objective dataset balancing with MLOPT-Ext [P6].

2.2.1 Localization of Facial Action Units

A subset of 30 AUs can be localized in the facial region, because they are related to the activation of specific facial muscles and tissues (Ekman and Friesen, 1978; Ekman et al., 2002). These AUs correspond to specific facial movements, such as raising the eyebrows or tightening the lips, which are the result of muscle contractions in well-defined regions of the face. A description of the facial muscles involved in the AUs is presented in Table 2.2. For example, AU 12 (lip corner puller) is associated to the zygomaticus major muscle, while AU 4 (brow lowerer) involves the corrugator supercilii muscle. This localization information is integrated into the domain-specific evaluation method RelEx for interpreting model predictions, presented in Chapter 8. The remaining AUs describe gross head and eye movements and are not considered in this localization-based analysis.

2.2.2 Co-Occurrence of Facial Action Units

The co-occurrence of AUs is driven by the anatomical structure of the face, as well as their relationship to FEs. Facial muscles often share tissue or are interconnected, resulting in the simultaneous activation of multiple AUs (Ekman and Friesen, 1978). Furthermore, specific combinations of AUs are indicative of FEs such as basic emotions and pain (Friesen, Ekman, et al., 1983; Craig et al., 1992). The role of AUs in the representation of FEs will be discussed further in the following section.

AUs can occur either in isolation or in combinations. The appearance of the AU combinations can be either additive or non-additive due to the anatomical structure of the face and the nature of muscle interactions (Ekman et al., 2002). In an additive combination, each AU retains its independent effect on facial appearance, meaning that the activation of one muscle group does not interfere with the activation of another. For example, AU6 (cheek raiser) and AU12 (lip corner puller) often appear together in a smile, but their individual effects remain visually distinct. In contrast, non-additive combinations occur when the activation of one AU changes or limits the appearance of another AU. For instance, in the combination by AU1 and AU4, the lifting of the inner brow of AU1 partially counteracts the lowering effect of AU4, limiting its movement to the outer brow area. These subtle variations require the expertise of certified FACS coders for accurate AU annotation.

The strength of the relationship between AUs depends on their anatomical proximity and how often they co-occur in FEs. From an affective computing perspective, the presence of multiple AUs in a single facial image introduces a multi-label classification problem. This domain knowledge about AU co-occurrences is integrated into the correlation-informed model training method, CorrLoss, discussed in Chapter 6.

Tab. 2.4.: Logical rules for mapping AUs to basic emotions (Friesen, Ekman, et al., 1983) and pain (Kunz and Lautenbacher, 2014). Adapted from [P7] (Rieger et al., 2025).

Facial sion	expres-	AU rule	Description of rule
Anger		$(4 \wedge 7) \vee (4 \wedge 5) \vee (4 \wedge 5 \wedge 7) \vee (17 \wedge 24)$	Lowered brows with tightened eyelids and, in some variants, additional tension in the lower face due to compressed lips and a wrinkled chin.
Fear		$1 \wedge 2 \wedge 4$	Raised inner and outer brows combined with brow tension and a furrowed brow.
Disgust		$9 \wedge 15 \wedge 16$	Wrinkled nose with downturned mouth corners and a lowered lower lip.
Happiness		$(6 \wedge 12) \vee (7 \wedge 12)$	Raised cheeks and pulled lip corners, corresponding to a Duchenne or non-Duchenne smile.
Sadness		$(1 \wedge 4) \vee (6 \wedge 15) \vee (11 \wedge 15) \vee (11 \wedge 17)$	Combinations of raised inner brows, lowered brows, and drooping mouth corners, sometimes with tension around the nose and chin.
Surprise		$(1 \wedge 2 \wedge 5) \vee (1 \wedge 2 \wedge 26)$	Raised eyebrows with widened eyes and either raised upper eyelids or a dropped jaw.
Pain		$(4 \wedge 6 \wedge 7 \wedge 9 \wedge 10) \vee (6 \wedge 7 \wedge 25 \wedge 26 \wedge 27) \vee (1 \wedge 2 \wedge 14) \vee (4 \wedge 6 \wedge 7 \wedge 14)$	Combinations corresponding to the four prototypical <i>faces of pain</i> : (1) narrowed eyes with furrowed brows and a wrinkled nose, (2) opened mouth with narrowed eyes, (3) raised brows with tension around the eyes, and (4) furrowed brows with narrowed eyes.

2.2.3 Hierarchical Relationship Between Facial Expressions and Facial Action Units

The hierarchical relationship between AUs and FEs provides insight into how expressions are formed from distinct and combined muscle movements (Tian et al., 2001). AUs serve as the building blocks, with specific combinations of AUs corresponding to FEs (Friesen, Ekman, et al., 1983; Kunz and Lautenbacher, 2014).

Table 2.4 outlines the rules used in this thesis for mapping AUs to basic emotions and pain expressions. The rules for basic emotions are grounded in the psychological studies by Friesen, Ekman, et al. (1983), who formalized these relationships through the Emotion Facial Action Coding System (EMFACS). For pain expressions, the rules are derived from psychological studies by Kunz and Lautenbacher (2014), who identified four distinct facial activity patterns based on AUs during pain expressions using cluster analysis. They also identified a stoic cluster with no facial patterns,

which is not considered in this thesis. The mapping of AUs to basic emotions and pain can be expressed using Boolean algebra, where the activation states of AUs are represented as Boolean variables (True or False). The logical connectives AND ($\wedge$) and OR ($\vee$) represent conjunction and disjunction, respectively. For the proposition to be true, the AND operator requires all AUs in the condition to be active (True), and the OR operator requires that at least one of the possible conditions (FEs) of combined AUs is active (True). Table 2.4 also provides a brief non-formal description of the rules to facilitate understanding of the facial movements involved.

This understanding of the hierarchical relationship between AUs and FEs is utilized in Chapter 6 for training multi-task FMC-Nets, and in Chapter 7 for the ensemble approach. Furthermore, pain-related AUs are evaluated using RelEx in Chapter 8. Finally, the frequency of AU occurrences varies across FEs, leading to class imbalances in the datasets. To deal with class imbalances, the MLOPT algorithm is introduced in Chapter 5.

2.2.4 Intensity Scale and Temporal Modality of Facial Expressions

FEs show considerable variation in intensity, which is determined by the degree of contraction of the facial muscles. Stronger contractions result in more pronounced expressions, while weaker contractions result in more subtle displays.

In FACS (Ekman et al., 2002), the intensity of AUs is measured on a scale from A to E, where A is the lowest visible intensity and E is the most extreme muscle contraction. This grading system can provide a fine-grained understanding of FEs, distinguishing, for instance, between a slight frown and an intense scowl. Similarly, pain-related expressions also show variations in intensity, reflecting the level of discomfort experienced by an individual (Kunz et al., 2019). However, annotating AU intensity requires certified FACS coders and is time consuming and costly. As a result, few datasets are annotated with intensity information, limiting the detailed analysis of FE and pain intensity in current research.

In addition to intensity, FEs unfold over time, following a temporal trajectory that is often modeled by three distinct phases (Han et al., 2023). The onset phase marks the initial contraction of the facial muscles as the expression begins to form. The apex phase represents the peak intensity of the expression, providing the clearest signal. The offset phase occurs when the muscles relax and the expression fades. In the real-world, however, these phases cannot account for all the temporal dynamics of FEs (Dhall et al., 2012).

In this thesis, Chapter 4 investigates how sequential models that process across multiple time steps might be influenced by demographic biases. In addition, the

varying intensity of AUs and their effect on FEs are incorporated into the decision-making process through the use of fuzzy logic in the ensemble method in Chapter 7.

2.2.5 Subject Demographic Metadata

Subject demographic metadata refers to individual characteristics, such as age, gender, and ethnicity that help ensure data diversity and improve model performance, especially for underrepresented subpopulations (Nourivandi et al., 2024). Many face image datasets have an unbalanced representation of demographic groups (Karkkainen and Joo, 2021). Addressing these biases during model training and evaluation (Mehrabi et al., 2021) is particularly critical for high-stakes applications, such as healthcare and security, where demographic biases could lead to inequitable outcomes and reduce trust in AI systems.

In this thesis, Chapter 4 describes in detail the annotation of several FER benchmark datasets with demographic metadata. MLOPT-Ext, an extension of MLOPT designed to incorporate additional objectives such as demographic metadata into dataset balancing, is presented in Chapter 5.

2.3 Deep Neural Networks

Deep neural networks (DNNs) have become essential tools in computer vision (LeCun et al., 2015; Zhang et al., 2023a), excelling in tasks such as FER (Leong et al., 2023). The following sections provide a short overview of the central computer vision task of image classification and the DNN architectures applied in this thesis.

Table 2.5 summarizes the classification setups, target tasks, and neural network architectures used in each publication.

Tab. 2.5.: Overview of learning setups, target tasks, and network architectures used across the included publications. Abbrev.: Train = model training performed, ML = multi-label; MC = multi-class; MT = multi-task.

Paper	Train	ML	MC	MT	AU	Emotion	Architecture(s)
P1	✓	✓			✓		ResNet-50 + custom head
P2	✓	✓			✓		ResNet18-112; VGG-16
P3		✓	✓	✓	✓	✓	OpenFace (AU toolkit); ResNet-50-based FER model
P4	✓	✓			✓		VGG-16
P5	✓	✓			✓		ResNet18-112; VGG-16
P6	✓	✓			✓		ResNet18-112
P7	✓	✓	✓	✓	✓	✓	ResNet-50 + custom head

2.3.1 Learning Setup

The central machine learning task in this thesis is image classification. Both FER and AU detection are framed as supervised learning problems, where the goal is to learn a mapping from an input facial image to a set of predefined class labels (Goodfellow et al., 2016).

Two main paradigms are employed. FER, which involves assigning a single emotion from a set of mutually exclusive categories (e.g., happiness, sadness), is treated as a multi-class classification problem. In contrast, AU detection, where a single face can exhibit multiple simultaneous muscle movements, is formulated as a multi-label classification problem. Across all contributions, model training is guided by minimizing a binary or categorical cross-entropy loss function, optimized using standard gradient-based methods such as Adam (Kingma, 2014) with mini-batch learning. However, a straightforward application of this framework to FER is non-trivial: severe class imbalance and complex statistical co-occurrence patterns of AUs motivate the data balancing and regularization techniques that form the core methodological contributions of this thesis.

2.3.2 Network Architectures

The deep learning models developed in this thesis are built upon two well-established CNN architectures: Residual Networks (ResNets) (He et al., 2016) and VGG (Simonyan and Zisserman, 2015). These were chosen for their proven effectiveness in image classification and their suitability for transfer learning.

As summarized in Table 2.5, the general approach across the publications involves using standard backbones (e.g., ResNet-50, VGG-16) pre-trained on the ImageNet dataset (Deng et al., 2009) and fine-tuning them for the specific AU detection and FER tasks. In addition to these standard models, several contributions [P2, P5, P6] employ a parameter-reduced ResNet-18 variant (ResNet18-112), developed in prior work to optimize efficiency without sacrificing performance (Rieger et al., 2019).

2.4 Evaluation Framework

The contributions of this thesis are validated through a multi-faceted evaluation framework designed to assess both model robustness and interpretability. The overarching strategy involves comparing model performance under controlled conditions, with specific metrics and protocols chosen to reflect the unique challenges of each research question.

Table 2.6 provides an overview of the evaluation frameworks applied in each pub-

lication, distinguishing between robustness-focused and interpretability-focused evaluations. To ensure traceability and provide a comprehensive overview, Table 2.7 summarizes the key performance metrics employed in each publication. Table 3.1 in Chapter 3 provides a detailed overview of the datasets used in the respective publications.

Tab. 2.6.: Overview of evaluation frameworks used across the included publications. Abbrev.: W = within-dataset; X = cross-dataset; SD-CV = subject-dependent cross-validation; RelEx = quantified saliency maps.

Paper	Evaluation focus	Robustness evaluation	Interpretability evaluation
P1	AU detection + CorrLoss	W (SD-CV)	RelEx
P2	AU detection	W (SD-CV)	RelEx
P3	Bias analysis (AU & emotion)	X (stratified by demographic group)	–
P4	AU detection + CorrLoss	W+X (SD-CV)	–
P5	AU detection	W+X (SD-CV)	RelEx
P6	AU detection + MLOPT balancing	w/ vs. w/o balancing (underrep. AUs)	–
P7	Multi-task + expert ensemble	W+X (SD-CV)	Ensemble disagreement / calibration

2.4.1 Evaluating Model Robustness

To provide a stringent test of generalization, the primary evaluation protocol is subject-dependent k-fold cross-validation. This ensures that models are always tested on data from subjects unseen during training, providing a realistic measure of performance on new individuals. This protocol is applied in two key scenarios:

- **Within-Dataset Evaluation:** Assesses a model’s ability to generalize to new subjects from the same data distribution.
- **Cross-Dataset Evaluation:** Provides a more challenging test of robustness by training on one dataset (e.g., lab-controlled) and testing on another (e.g., in-the-wild), measuring resilience to domain shift, in other words, model generalization.

Model performance is quantified using a suite of standard metrics appropriate for the task. The F1-score is prioritized for most classification tasks due to its effectiveness in handling the class imbalance prevalent in affect recognition datasets. Other key metrics include AUROC for threshold-independent evaluation, alongside diagnostics like training/validation loss to monitor for overfitting.

Tab. 2.7.: Quantitative evaluation overview across the included publications: evaluated entity, quantitative aspect (one aspect per row), and corresponding metric(s). Abbrev.: PP = predictive performance; RelEx = quantified saliency maps.

Paper	Evaluated entity	Quantitative aspect	Metric(s)
P1	AU model	PP (CorrLoss)	F1-score
		RelEx	μ -score
P2	AU model	PP	F1-score
		RelEx	μ -score
P3	AU and emotion models	Bias analysis	F1-score
	Dataset metadata	Annotation quality	Krippendorff's α
P4	AU model	PP (CorrLoss)	F1-score; training/validation loss
P5	AU model	PP	F1-score
		RelEx	μ -score
P6	MLOPT outcome	Optimization behavior and outcome	Objective function convergence; resulting distribution (MeanIR)
	AU model	PP (balanced classes)	F1-score
P6	AU model	PP (balanced classes, MLOPT)	F1-score; objective function convergence
P7	Multi-task model	PP	Accuracy; AUROC; F1-score; precision; recall; confusion matrix
	Expert ensemble	Calibration	Calibration plot + ensemble disagreement

For data balancing contributions (Chapter 5), effectiveness is also measured using the Mean Imbalance Ratio (MeanIR) (Charte et al., 2015a). This metric quantifies overall class imbalance by averaging the imbalance ratios of each class, where a lower value indicates a more balanced dataset. It is defined as:

$$MeanIR = \frac{1}{n} \sum_{i=1}^n IRLbl(x_i), \quad (2.1)$$

where n is the number of classes, and $IRLbl(x_i)$ is the imbalance ratio for class i , computed as:

$$IRLbl(x_i) = \frac{\max(\mathbf{x})}{x_i}, \quad (2.2)$$

where x_i is the count of the i -th class, and $\max(\mathbf{x})$ represents the maximum value across all classes in the distribution $\mathbf{x}$.

Finally, to ensure the quality of the novel demographic annotations, inter-rater reliability was measured using Krippendorff's alpha (Krippendorff, 2018). This is a robust statistical measure that quantifies the level of agreement between multiple coders, correcting for the possibility of agreement occurring by chance. A value of $\alpha = 1$ indicates perfect agreement, while $\alpha = 0$ represents agreement no better than random chance. Values above 0.80 are generally considered to indicate high reliability and values above 0.67 sufficient for preliminary conclusions (Krippendorff, 2004). This metric was used for validating that the annotated metadata is reliable for the bias analyzes presented in this thesis (Chapter 4).

2.4.2 Evaluating Model Interpretability

The evaluation of interpretability in this thesis is performed through a computational and comparative framework, which intentionally omits subjective user studies to first establish an objective baseline. The quality of the two primary interpretability contributions, the fuzzy logic ensemble and the quantified saliency maps (RelEx), is assessed with distinct approaches tailored to each method.

For the fuzzy logic ensemble, interpretability is inherent in its human-understandable, rule-based design. Therefore, its evaluation focuses on two key aspects: its contribution to overall model performance and its ability to generate reliable disagreement-based risk signals. The quality of these estimates is validated using calibration plots, which measure how well a model's stated confidence aligns with its actual predictive accuracy.

For saliency-map-based explanations, this work uses standard methods like Gradient-

weighted Class Activation Mapping (Grad-CAM)(Selvaraju et al., 2017) and Layer-wise Relevance Propagation (LRP)(Bach et al., 2015):

- **Gradient-weighted Class Activation Mapping (Grad-CAM):** Uses the gradients of a target class flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the input for predicting that class (Selvaraju et al., 2017). This method is primarily designed for CNNs.
- **Layer-wise Relevance Propagation (LRP):** Explains a model's prediction by distributing the prediction score backward through the network layers, attributing relevance scores to each input feature (Bach et al., 2015). Unlike gradient-based methods, LRP aims to provide explanations of how each input pixel contributes to the final output, often resulting in crisper and more detailed relevance maps. This method can be applied to any feedforward neural network.

While these saliency map techniques provide valuable visual insights, they have been criticized. A significant body of research has raised valid concerns about their faithfulness and reliability (Adebayo et al., 2018). To address the limitations of subjective visual inspection and provide a more rigorous assessment, interpretability is assessed using the developed RelEx framework. RelEx offers a systematic, quantitative evaluation by measuring the alignment of these saliency maps with expert-defined anatomical regions, which provides a domain-specific measure of whether a model's reasoning is plausible.

2.4.3 Correlation Coefficients for Model Regularization

A core methodological contribution of this thesis is a regularization term based on statistical correlation (Chapter 6). This approach utilizes two key metrics to measure the relationship between model predictions and target labels:

- The Pearson Correlation Coefficient (PCC) quantifies the strength and direction of a linear relationship between two variables. It is used to ensure that the model learns the general associative patterns between classes.
- The Concordance Correlation Coefficient (CCC) evaluates the agreement between two variables by measuring how much they deviate from a 45° line of perfect agreement. Unlike PCC, CCC penalizes both low correlation and differences in mean and variance. It is used to enforce a stricter, more precise alignment between the predicted and target co-occurrence patterns.

By applying both PCC and CCC, the proposed regularization method can be tuned to enforce either general statistical relationships or strict predictive agreement (AU–AU, AU–FE), providing a flexible mechanism for integrating domain knowledge.

Datasets and Data Preprocessing

In the context of this thesis, a variety of benchmark datasets were collected and preprocessed. The selected datasets were chosen due to their prominence in benchmarking FE and AU recognition tasks. The following sections provide an overview of the data acquisition process, dataset characteristics (see Table 3.1), and the preprocessing steps taken to prepare the data for deep learning model development.

3.1 Dataset Acquisition and Management

The datasets used in this work have been carefully acquired from multiple sources, taking advantage of the collaborations and resources made available through the affiliation with Fraunhofer IIS. For instance, the *Actor Study Database* (Seuss et al., 2019) was collected internally by Fraunhofer IIS, the *Aff-Wild2 Database* (Zafeiriou et al., 2017) was obtained through participation in the ABAW challenge (Kollias et al., 2020), and the *ERIK dataset* was provided by the ERIK research project. All datasets were obtained under research licenses. The handling of the data ensured compliance with the licensing terms and data protection regulations, including GDPR. Handling of sensitive subject data was carefully managed to maintain confidentiality and privacy, with access and backup restricted by secure storage protocols. Additionally, the publication of images from these datasets is restricted, reflecting the sensitivity of the data. This approach ensured a responsible and structured method of dataset aggregation and management throughout the research process.

3.2 Dataset Characteristics

The datasets used in this thesis span a variety of recording setups and coding schemes. Table 3.1 summarizes the key characteristics of the datasets that may affect the research results. Additionally, the publication IDs of my publications where these datasets have been used are listed. The number of subjects (*subj.*) determines the variety and generalizability of models and analyzes. The total number of frames (*frames*) reflects the size of the dataset, where small datasets may limit the gen-

Tab. 3.1.: Overview of datasets used in this thesis and corresponding publications in which they were employed. Adapted from Pahl et al. (2022), Table 1.

Dataset	Papers	Subj.	Frames	Setup / Affect	Affect type	AUs	FEs / other labels
Actor Study	P1, P2, P4–P7	21	146 663	lab, acted / spont.	emotion-related	1, 2, 4–7, 9–18, 20, 22–27, 38, 43, 45, 46 (<i>frame</i>)	amusement, anger, boredom, contempt, disgust, enjoy, fear, happiness, interest, pride, relief, sadness, surprise (<i>sequence</i>)
AffWild2	P3, P4, P6, P7	N/A	2 395 718	wild, spont.	emotion-related, dimensional	1, 2, 4, 6, 12, 15, 20, 25 (<i>frame</i>)	anger, disgust, fear, happiness, neutral, sadness, surprise; valence, arousal (<i>frame</i>)
BP4D	P1, P3, P4, P7	41	366 954	lab, spont.	emotion-related, pain	1, 2, 4–7, 9–20, 22–24, 27–39 (<i>frame</i>)	anger, disgust, embarrassment, fear, happiness, pain, sadness, surprise (<i>sequence</i>)
CK+	P2, P3, P4, P5, P6, P7	123	10 733	lab, acted	emotion-related	1, 2, 4–7, 9–18, 20–31, 34, 38, 39, 41–46 (<i>sequence</i>)	anger, contempt, disgust, fear, happiness, sadness, surprise (<i>sequence</i>)
DISFA	P3	27	130 815	lab, spont.	emotion-related	1, 2, 4–6, 9, 12, 15, 17, 20, 25, 26 (<i>frame</i>)	–
EmotionNet (Man.)	P3, P4, P6	N/A	24 598	wild, spont.	emotion-related	1, 2, 4–6, 9, 10, 12, 15, 17, 18, 20, 24–26, 28, 43 (<i>frame</i>)	–
ERIK	P3	13	2 700	lab, acted	emotion-related	1, 2, 4–7, 9–18, 20, 22, 24–38, 43, 45, 51–58, 61–64 (<i>frame</i>)	anger, anxiety, neutral, joy, sadness, surprise (<i>frame</i>)
GFT	P3, P4	96	172 800	lab, spont.	emotion-related	1, 2, 4–7, 9–12, 14, 15, 17–19, 22–24, 28 (<i>frame</i>)	–
UNBC	P2, P3, P5	25	48 398	lab, spont.	pain-related	4, 6, 7, 9, 10, 12, 20, 25–27, 43 (<i>frame</i>)	pain (<i>frame</i>)

eralizability of models and analyzes. The environment (*lab* vs. *wild*) affects the variability and realism of the data, with lab settings providing controlled conditions and in-the-wild data capturing more natural and diverse scenarios. The affect type (*spont.* vs. *acted*) emphasizes the authenticity of the expressions, which is critical for ensuring real-world relevance. Finally, the annotation level (*seq.* vs. *frame*) defines the granularity of the labels, whether at the frame level or broader trends across sequences. Note that information about demographic metadata information is summarized in Chapter 4. The different datasets serve as training and test datasets for methods and models in the experiments performed in this thesis. The following dataset summaries, including dataset statistics, highlight key details relevant to this research; additional information can be found in the referenced papers.

Actor Study Database

This dataset contains recordings of 21 actors performing both spontaneous and acted FEs under controlled laboratory conditions. High-resolution cameras captured the images in RGB format at a resolution of 1624×1236 pixels, ensuring uniform illumination and minimal noise. After preprocessing (described in Section 3.3), cropped face regions have an average size of 351×258 pixels with a standard deviation of 23×15 pixels, providing relatively uniform input. Actors participated in tasks that elicited both single and combined AUs and responses to emotional scenarios. It contains 777 video sequences (68 minutes in total) with frame-based AU intensity annotations (three levels) and sequence-based FE annotations for 13 expressions. This dataset has been partially released. A detailed description can be found at Seuss et al. (2019).

Aff-Wild2 Database

This dataset is an in-the-wild dataset sourced from YouTube, with a wide range of video qualities, head poses, lighting conditions and resolutions, making it suitable for real-world affect recognition. The dataset contains 303,250 frames from 47 subjects, annotated frame by frame for AUs and FEs. The resolution varies, with an average of 1030×630 pixels, while cropped face regions have an average size of 266×196 pixels with a standard deviation of 173×131 pixels, reflecting the variability of video sources. Not all frames are annotated for each class, adding complexity to model training. Aff-Wild2 is a widely used benchmark, often used in challenges such as CVPR's Affective Behavior Analysis in the Wild (ABAW) challenge (Kollias et al., 2023). More details can be found in Kollias et al. (2019), Kollias and Zafeiriou (2019), and Zafeiriou et al. (2017).

BP4D-Spontaneous Dataset

This dataset captures natural, spontaneous FEs in both 2D and 3D formats, using synchronized RGB cameras and depth sensors. It comprises 328 videos from 41 subjects and is specifically designed to elicit eight target FEs through various emotional stimuli, including video clips, startling probes, threats, cold pressor tasks (dipping a hand in ice water), and smells. High-resolution images have an average resolution of 1040×1392 pixels, while cropped face regions have an average resolution of 782×585 pixels with a standard deviation of 45×31 pixels, ensuring consistency for facial analysis. The dataset serves as a benchmark for spontaneous AU detection, providing frame-level AU annotations and sequence-level FE labels. Intensity annotations are included for a subset of five AUs, providing finer granularity for analysis. The extended version, BP4D+, includes additional subjects and multimodal data,

further enhancing its utility for affective computing research. More information is available at Zhang et al. (2014) and Zhang et al. (2016).

Extended Cohn-Kanade (CK+) Database

This dataset contains 593 video sequences from 123 subjects, capturing posed FEs of basic emotions along with contempt. Each sequence begins with a neutral expression and progresses to a peak emotional expression. The images, captured in controlled laboratory settings, are available in both RGB and grayscale formats, with an average resolution of 640×490 pixels. Cropped face regions have an average size of 296×226 pixels with a standard deviation of 28×23 pixels, ensuring consistency in face analysis. The dataset provides frame-level AU annotations and sequence-level FE labels and was one of the first open-source datasets in this domain, making it a widely used benchmark for the analysis of acted FEs. Detailed information can be found at Lucey et al. (2010).

Denver Intensity of Spontaneous Facial Action (DISFA) Database

This dataset contains high-resolution RGB videos of 27 participants recorded under controlled conditions with uniform lighting and a blue background to ensure consistency. Each subject watched a 4-minute video clip designed to elicit spontaneous FEs. The videos have an average resolution of 1024×768 pixels, while the cropped face regions have an average size of 372×272 pixels with a standard deviation of 31×26 pixels, ensuring consistency for face analysis. The dataset is annotated frame by frame for 12 AUs, with intensity ratings provided on a five-point scale. The extended version, DISFA+, introduces additional subjects. This dataset is widely used as a benchmark for AU intensity detection due to its rich annotation and high quality images. For more details, see Mavadati et al. (2013) and Mavadati et al. (2016).

EmotioNet

This dataset is a large-scale collection designed to capture a diverse range of FEs in real-world settings. It consists of over one million images collected from the web using emotion-related keywords. While most of the dataset is automatically annotated using a machine learning-based approach for AUs, this work focuses on the manually annotated subset of 24,598 images, which provides higher reliability. The images in the manually annotated subset have an average resolution of 543×584 pixels, while the cropped face regions have an average size of 318×238 pixels with a standard deviation of 248×188 pixels, reflecting the variability in image quality and source conditions. This subset provides frame-level AU annotations and is particularly valuable for benchmarking affective computing models due to its comprehensive coverage of facial expressions and high annotation quality. For more information,

visit Fabian Benitez-Quiroz et al. (2016).

ERIK

This dataset consists of high-resolution RGB images of 13 children collected under controlled laboratory conditions with optimal lighting and uniform backgrounds. The images have an average resolution of 460×460 pixels, while cropped face regions average 295×220 pixels with a standard deviation of 23×12 pixels, ensuring consistency for face analysis. The dataset focuses on acted FEs and is annotated frame by frame for AUs and FEs. It emphasizes consistency in lighting and imaging conditions, providing high-quality data suitable for detailed facial analysis. While the dataset remains unpublished, it serves as a valuable resource for studying facial expressions in children, a population often underrepresented in affective computing research. More details can be found in the project documentation.¹

Sayette Group Formation Task Spontaneous Facial Expression (GFT) Database

This dataset captures spontaneous FEs during social interactions. It contains RGB videos of 96 participants interacting in groups of three. The videos have an average resolution of 720×480 pixels, while cropped facial regions average 266×222 pixels with a standard deviation of 44×39 pixels, ensuring consistency for facial analysis. The dataset uniquely incorporates social dynamics, with some participants consuming alcoholic beverages to elicit more naturalistic and varied FEs. Although recorded in a laboratory, the setup is less controlled than other datasets, with challenges such as occlusion from glasses or other participants, as well as varying lighting conditions. The videos are annotated frame by frame for AUs, providing detailed information about spontaneous facial expressions in a group context. This dataset is particularly valuable for studying facial behavior in social settings, where interactions and dynamics play an important role. For more details, see Girard et al. (2017).

UNBC-McMaster Shoulder Pain Expression Archive Database

This dataset focuses on facial expressions associated with pain. It contains video sequences of 25 subjects experiencing shoulder pain, recorded under controlled laboratory conditions. The videos have an average resolution of 341×240 pixels, while cropped facial regions have an average size of 165×128 pixels with a standard deviation of 15×12 pixels, reflecting the lower image quality compared to other datasets. The dataset is annotated frame by frame for AUs and includes pain intensity ratings, making it a valuable resource for studying pain-related facial expressions. The subjects are predominantly older adults, providing a unique perspective compared to datasets that typically include younger participants. Despite its lower resolution,

¹<https://www.interaktive-technologien.de/projekte/erik>

the dataset is still widely used as a benchmark for FER in medical contexts. More details can be found at Lucey et al. (2011).

These datasets exhibit variation in experimental setups, coding schemes, and the inclusion of spontaneous versus acted AUs and FEs. This diversity underscores the challenge of achieving model robustness.

3.3 Data Preprocessing

Data preprocessing played an important role in preparing the datasets for deep learning models in this thesis. While standard steps such as image resizing and pixel normalization were used, as detailed in my publications, this section highlights two challenging aspects of facial image data preprocessing: dataset aggregation and face detection.

STRAND Framework for Dataset Aggregation: To address the heterogeneity of dataset annotations, the *Scripted Tools (for) Reading And Normalizing Databases* (STRAND) framework was developed in collaboration with colleagues at Fraunhofer IIS. It establishes a standardized schema for integrating multiple, diverse datasets, ensuring consistency while preserving the integrity of original annotations. The schema uses key-value pairs with predefined types that facilitate the standardization of inconsistent annotations, such as mapping terms like "happy" and "happiness" to a unified label. Within this schema, each key corresponds to a column, and each row represents a single image or video frame. In addition to unifying class annotations, the STRAND framework integrates critical metadata, including demographic information (discussed in Section 4), image paths, and licensing details. By creating a single, consistent data-access pipeline, the framework eliminated the need to revisit the specific annotation schema of each individual dataset for every new research approach, thereby ensuring methodological consistency and accelerating the research lifecycle.

The framework also provides pipelines for key preprocessing tasks, such as automatic face detection using SHORE™ (Küblbeck and Ernst, 2006) or OpenCV (Bradski, 2000), detection of 68 facial landmarks with SHORE™, and the automatic annotation of unlabeled data using pre-trained models. These pre-calculations accelerated the subsequent model training and evaluation process further. The framework supports data export in multiple formats, such as Parquet, CSV, JSON, and `pandas.DataFrame`, ensuring flexibility for various research needs. A structured workflow was established for incorporating new datasets, with predefined templates and review cycles ensuring consistency and efficiency. Although the STRAND framework itself is not

published due to Fraunhofer regulations (see Section 1.5.2), it served as a cornerstone of the data preprocessing efforts in this research. Its implementation provided a standardized and efficient approach to aggregating datasets, streamlining the integration of smaller datasets with varied annotations, and improving the overall quality of data preparation.

Face detection: Face detection was another critical aspect of preprocessing. All models developed in this thesis were trained on cropped face images. Two main tools were used for automatic face detection: SHORE™, a sophisticated high-speed object recognition software developed by Fraunhofer IIS (Küblbeck and Ernst, 2006), and the OpenCV DNN module, a well-known open source framework (Bradski, 2000). Both tools were able to detect non-frontal faces. The choice of face detection method varied between the corresponding publications. For publications P5 and P6, SHORE™ was used, while OpenCV was used for P1, P2, P4, and P7. For publication P3, face detection methods specific to the models evaluated were used. Special considerations were required for the Aff-Wild2 dataset, which contained multiple faces in some frames. In these cases, manual annotation of face regions was performed by students prior to automatic face detection according to the descriptions in the annotations. When the Aff-Wild2 annotations did not specify the target face (e.g., left or right), the largest face in the frame was selected to maintain consistency. To automate this process, a browser-based tool called *Videotator* was developed. These preprocessing efforts ensured the extraction of high quality face data.

Bias Analysis in Datasets and Models [P3]

This chapter addresses *RQ1: How can fairness across demographic groups be measured in facial expression recognition datasets and trained models?*

The scientific contributions, visualizations, and results presented in this chapter are derived from my publication [P3] (Pahl et al., 2022).

This chapter details the annotation of demographic metadata and analyzes demographic imbalances in facial expression recognition datasets, along with their impact on model performance and fairness.

This chapter creates domain knowledge in the form of subject demographic metadata (Section 2.2.5).

4.1 Motivation and Related Work

One of the major challenges in facial expression recognition is the uneven performance of models across different demographic groups (Nourivandi et al., 2024). This selection bias occurs when training datasets are not representative of the broader population, which is often the case for datasets of facial images (Karkkainen and Joo, 2021), resulting in models that perform well on overrepresented groups but poorly on underrepresented ones. Such model biases not only compromise accuracy, but also raise ethical concerns about the use of these systems in real-world settings. For instance, the Google Lens incident, where images of people of color were labeled as gorillas, led to the removal of the label entirely (Vincent, 2018). Similarly, Amazon's AI recruiting tool was found to disadvantage women, highlighting the broader societal implications of biased AI systems (Dastin, 2018). While there are few fair datasets with demographic metadata such as FairFace (Karkkainen and Joo, 2021) or the Balanced Faces in the Wild (Robinson et al., 2023), they are not annotated for facial expression recognition tasks, limiting their utility in this domain.

Addressing demographic bias in facial expression recognition is challenging, because demographic metadata annotations such as age, gender, and ethnicity are often incomplete or missing in available datasets. As shown in Table 4.1, only a few of the

Tab. 4.1.: Overview of datasets and available demographic metadata for gender, ethnicity, and age. Adapted from [P3] (Pahl et al., 2022).

Dataset	Per Subject	Statistics	Annotations
Actor Study	-	✓	AUs, FEs
Aff-Wild2	-	-	AUs, FEs
BP4D / BP4D+	gender	✓	AUs, FEs
CK+	-	✓	AUs, FEs
DISFA / DISFA+	-	✓	AUs
ERIK	(forenames)	-	AUs, FEs
EmotioNet (Man.)	-	-	AUs
GFT	✓	✓	AUs
UNBC	-	gender	AUs, FEs

AU and FE benchmark datasets used in this thesis, such as BP4D and GFT, include demographic metadata per subject, while others, particularly in-the-wild datasets such as Aff-Wild2 and EmotioNet, do not. On the other hand, statistics such as the distribution for gender and ethnicity or an age range are more often disclosed. An overview of other dataset characteristics can be found in Section 3.2. Without demographic annotations, it becomes difficult to identify and correct biases during model training and evaluation.

To address this limitation, a contribution of this thesis is the post-annotation of demographic metadata for each subject in these benchmark datasets.

4.2 Demographic Metadata Annotation

The missing demographic metadata for each subject in the datasets listed in Table 4.1 was retrospectively annotated, with the exception of the Actor Study. The Actor Study was excluded because the ground truth demographic annotations were reserved for future publication. For the Aff-Wild2 dataset, only AU annotated sequences were included in the demographic annotation process. Furthermore, the dataset was split into its official subsets, AU validation (AU-V) and AU training (AU-T).

Demographic metadata annotation was carried out by four independent coders, who reviewed images or videos of the subjects to estimate gender (male or female), age (as an integer), and ethnicity based on appearance. The ethnic categories were defined as African-American, Asian, Euro-American, or Hispanic, consistent with those commonly used in the demographic metadata statistics. Additionally, the coders annotated the presence of eyeglasses and facial hair (beard). Coders were given access to statistical distributions of ethnicity, gender, and age in the datasets

to guide their decisions and ensure consistency.

To quantify the reliability of the demographic annotations, Krippendorff's α (Krippendorff, 2018) is calculated for gender, ethnicity, and age across all datasets (see Section 2.4.1). For the metadata annotations, the average Krippendorff's α across all datasets and attributes is 0.87, indicating a high level of inter-rater agreement. Gender annotations show the highest reliability per dataset, while age estimates show the lowest agreement, which may be due to the annotation in integers. To account for this variability, age data are grouped into ten-year intervals for analysis. Overall, the demographic metadata annotations are found to be sufficiently reliable for use in the bias analysis of datasets (Section 4.3.1) and trained models (Section 4.3.2), as well as for the MLOPT-Ext method introduced in Section 5 that balances metadata during data preprocessing.

4.3 Summary of Key Results

Using the newly annotated metadata, the analysis in [P3] (Pahl et al., 2022) yields two main findings: (1) major demographic imbalances in FER/AU benchmark datasets, and (2) corresponding performance disparities of state-of-the-art models across demographic groups.

Unless stated otherwise, learning setups, architectures, evaluation frameworks, metrics, and datasets follow Tables 2.5–3.1.

4.3.1 Bias Analysis in Datasets [P3]

The analysis of the demographic bias in the annotated datasets reveals large imbalances, particularly in the distributions of age and ethnicity. The combined annotated data consists of 1,070,585 frames from 526 subjects.

Key observations:

- **Age skew.** ~71% of the frames contain subjects between the ages 20–30. While datasets such as ERIK (children) and UNBC (subjects > 35 years) help to broaden the age range, they remain underrepresented (Fig. 4.1).
- **Ethnicity imbalance.** Euro-American dominates (62.08%), with lower shares for Asian (15.97%), African-American (11.13%), and Hispanic (10.65%) groups (Fig. 4.2).
- **Gender near-parity.** But within specific ethnic groups, gender representation varies (e.g., fewer female African-American frames; comparatively more female Asian and Hispanic frames).

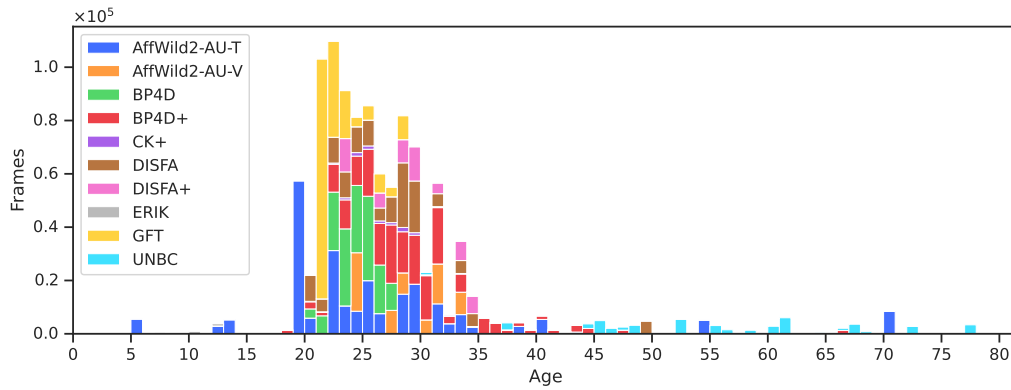


Fig. 4.1.: Age distribution of all combined datasets, with hue for dataset. From [P3] (Pahl et al., 2022).

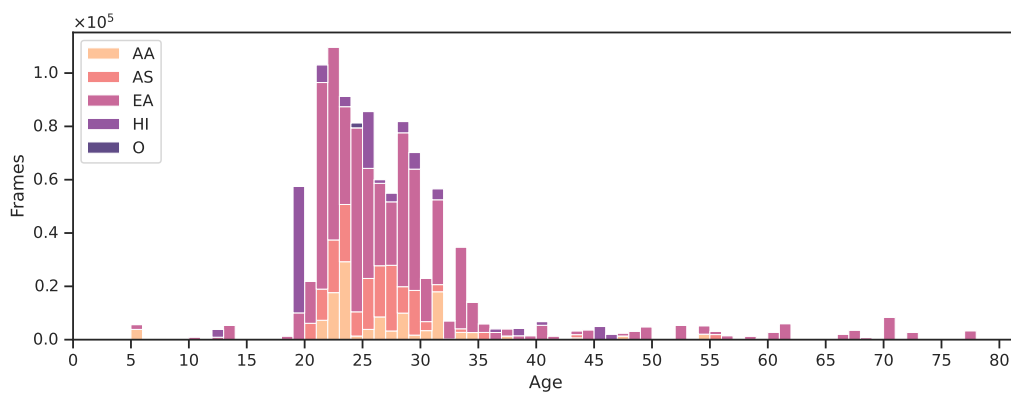


Fig. 4.2.: Age distribution of all combined datasets, hue ethnicity (AA: African-American, AS: Asian, EA: Euro-American, HI: Hispanic, O: Other). From [P3] (Pahl et al., 2022).

- **Occlusions are underrepresented.** Only 11.63% of frames show facial occlusions such as glasses; 30.03% of subjects have facial hair.

4.3.2 Bias Analysis in Models [P3]

The analysis of model performance across demographic groups for multiple AUs and basic emotions identified several biases. Two models were evaluated: NISL2020 (Deng et al., 2020), which combines multi-task CNN with RNNs and achieved first place in AU detection and third place in FER in the *Affective Behavior Analysis in-the-wild* competition (Kollias et al., 2020); and OpenFace 2.0 (Baltrusaitis et al., 2018), a widely used real-time, open-source toolkit that uses SVMs and CNNs. The models were evaluated only on datasets that were not used for model training. Key observations:

- **Age effects.** Best AU detection observed for ages 15–34. There is a noticeable drop in performance for older adults, particularly those over 55 years of age. This is consistent with the dataset age skew.
- **Ethnicity effects.** For AUs, NISL2020 performs best on Asian subjects and weakest on African-American subjects (especially women). For FEs, NISL2020 is strongest on African-American subjects and weakest on Hispanic subjects. OpenFace AU performance is strongest on African-American subjects—despite Euro-American dominance in the data.
- **Gender effects.** Mild but present: OpenFace favors male faces for AU detection, whereas NISL2020 favors female faces.
- **Occlusion sensitivity.** Both systems degrade with eyeglasses. AU detection performance related to movements near the eyes, such as AU 4 (brow lowerer) or AU 6 (cheek raiser) is particularly affected.

4.4 Summary and Discussion

The research question **RQ1**: *How can fairness across demographic groups be measured in facial expression recognition datasets and trained models?* was addressed in this chapter by systematically annotating demographic metadata and analyzing demographic biases in several benchmark datasets and state-of-the-art models. A key novel contribution of this study is the annotation of previously missing demographic metadata across multiple benchmark datasets. While institutional policies currently restrict the public release of these annotations (see also Section 1.5.2), the process and findings detailed here provide a blueprint for the field.

Substantial dataset imbalances lead to fairness effects: The analysis reinforced the reality that the benchmark datasets used in this field are deeply imbalanced, with a stark skew towards younger, Euro-American subjects. Furthermore, the subsequent model analysis revealed that this data-level bias translates directly into decreased model performance. The sharp decline in accuracy for older adults and certain ethnic minorities is a direct consequence of unrepresentative training data. This finding empirically establishes the connection between a lack of data diversity and model robustness and fairness failures.

Analysis motivates bias mitigation: By quantifying these demographic distributions, this chapter provides the foundations for the bias mitigation strategies that follow. Knowing the distribution of demographic attributes enables the implementation of bias mitigation strategies, such as balancing datasets or raising awareness of potential model biases. This analysis, therefore, serves as the direct motivation

and empirical foundation for RQ2, which tackles the challenge of algorithmically rebalancing these datasets, as will be detailed in Chapter 5.

Multi-Label Data Balancing: MLOPT and MLOPT-Ext [P6]

This chapter addresses *RQ2: How can multi-label datasets be balanced across multiple objectives in a way that remains suitable for training deep neural networks?*

The scientific contributions, visualizations, and results discussed in this chapter are derived from my publication [P6] (Rieger et al., 2020b). This contribution is extended by additional results (Appendix A).

This chapter details the development of the *Multi-Label Balancing Optimizer (MLOPT)*. MLOPT reframes multi-label resampling over unique label-sets as a bounded non-linear optimization problem. The core method was introduced in publication ([P6] (Rieger et al., 2020b)). This chapter synthesizes and extends the original contribution by: (1) providing a comprehensive literature review, which was not possible in the original short paper, to contextualize both the core method and the new experimental results, (2) presenting a unified formulation of the algorithm, which incorporates a multi-objective extension for demographic attributes, MLOPT-Ext, and (3) detailing new, unpublished experimental results that further validate MLOPT-(Ext) (see Appendix A). This chapter also reports empirical effects on imbalance metrics and downstream model performance.

This chapter incorporates domain knowledge in the form of subject demographic metadata (Section 2.2.5), whose creation was detailed in the previous Chapter 4.

5.1 Motivation and Related Work

For real-world applications, imbalanced datasets are a prevalent issue, presenting critical challenges to the development and performance of AI systems (Giorgio et al., 2023) (see previous Chapter 4). For example, in fraud detection, addressing class imbalance is essential to prevent models from overlooking infrequent but financially critical fraudulent activity (Gupta et al., 2023). Similarly, in object detection, models trained on imbalanced datasets tend to prioritize frequent objects, such as cars or pedestrians, while often neglecting rare objects such as traffic lights, which can lead

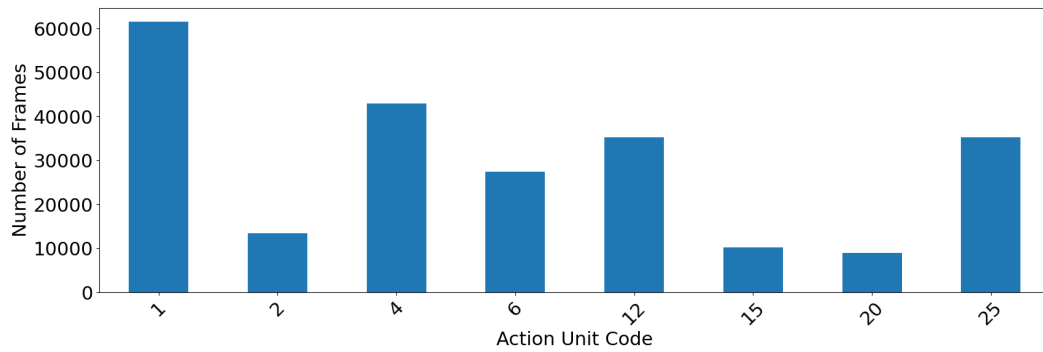


Fig. 5.1.: Distribution of AUs in a merged database comprising Actor Study, Aff-Wild2, CK+, and EmotioNet (Manual).

to problems in real-world applications (Oksuz et al., 2021).

The growing body of research on imbalanced learning reflects the increasing importance of addressing this topic (Rezvani and Wang, 2023). This concern is particularly relevant in FER, where sampling biases in datasets can lead to unfair models, as discussed in Section 4 based on [P3] (Pahl et al., 2022). In addition to equity across demographic groups, class imbalance is a persistent problem in FE and AU datasets (Nourivandi et al., 2024). The class imbalance in multi-label AU datasets typically results from two main factors: (1) the unequal frequency of AUs across different FEs, due to the natural occurrence rates of certain FEs and variations in video length, and (2) inconsistent subsets of annotated AUs across datasets, largely due to the labor-intensive and costly nature of the annotation process. Figure 5.1 shows the imbalanced distribution of AU classes in a merged dataset used in this approach. Addressing these imbalances is essential for developing models capable of accurately interpreting FEs and AUs across different classes and demographic groups.

Imbalanced datasets can be addressed by improved training strategies, but also by data *pre-processing techniques* (Rezvani and Wang, 2023). Choosing data preprocessing techniques has the advantage of being model agnostic and can be integrated into existing pipelines and workflows without overhauling the entire training process. Balancing algorithms designed for single-label problems with only one label per sample have advanced greatly (Giorgio et al., 2023; Rezvani and Wang, 2023; Chawla et al., 2002), but balancing multiple classes for one sample as found in *multi-label balancing problems* remains underexplored (Rezvani and Wang, 2023). This leaves a gap in effectively addressing the complexities of multi-label datasets, where multiple classes may exist simultaneously in a data sample. Balancing multiple classes for multi-label samples presents a unique challenge, as changing the frequency of one class may inadvertently change the frequency of others. This problem requires

consideration not only from the perspective of imbalance among labels, but also among unique label combinations, so called label-sets.

Existing multi-label balancing algorithms can be broadly categorized based on their approach to *oversampling* and *undersampling*. Oversampling-based methods clone data samples until a sufficient balance is reached, while undersampling methods delete samples. Existing approaches such as LPROS and LPRUS (Charte et al., 2013) perform over- and undersampling, respectively, by using *label-sets* for the initial step, followed by random sampling. ML-ROS and ML-RUS (Charte et al., 2015a) focus on individual labels rather than label-sets, using the MeanIR metric (Section 2.4.1) to identify minority labels. Oversampling techniques are generally better suited for deep learning models that require large datasets. Unlike undersampling, oversampling helps to augment minority classes without losing valuable information. However, excessive cloning can lead to overfitting in a DNN, where the model can become, despite introduced augmentation, too closely fitted to these data points. The MLOPT algorithm [P6] (Rieger et al., 2020b) addresses this problem by oversampling in a controlled manner to provide a method that is suitable for deep learning approaches.

In contrast, there are also methods that create *synthetic samples*. The REMEDIAL algorithm (REsampling MultilabEl datasets by Decoupling highly ImbAlanced Labels) (Charte et al., 2019) pre-processes datasets by decoupling minority or majority labels from samples to facilitate subsequent balancing steps, deliberately modifying labels for specific instances. Another prominent method is Multi-Label SMOTE (MLSMOTE) (Charte et al., 2015b), which extends SMOTE (Sivakumar et al., 2022) to generate synthetic samples for minority classes using nearest neighbor interpolation. This adaptation preserves the relationships between multiple labels in multi-label classification tasks. While synthetic samples and labels can work well to address imbalances in datasets, they are not universally applicable or beneficial for every use case. In some scenarios, generating synthetic data can introduce artifacts or patterns that do not accurately reflect the underlying characteristics of the real data. For example, in nuanced domains such as medical diagnosis or affect recognition, the subtleties of real human expressions or medical symptoms may be difficult to accurately replicate through synthetic means. To ensure applicability in such nuanced domains and to minimize additional noise, the MLOPT [P6] (Rieger et al., 2020b) algorithm does not generate synthetic labels or samples.

Existing over- and undersampling algorithms such as LPROS, LPRUS, ML-ROS, and ML-RUS focus primarily on class distributions without considering additional objectives, such as fairness across demographic groups. Similarly, synthetic methods such as REMEDIAL and MLSMOTE preserv relationships between labels but neglect demographic metadata. These approaches are effective for mitigating class imbalance but

Method	Approach	Bound	Synthetic
REMEDIAL	Decoupling of labels	Dataset size	Labels
MLSMOTE	Interpolating nearest neighbors	Imbalance metric	Images and labels
ML-ROS	Random OS of minority labels	Dataset size	No
LPROS	Random OS of minority label-sets	Dataset size	No
MLOPT	Optimized OS of label-sets	Label-set frequency	No

Tab. 5.1.: Overview of competitive multi-label oversampling (OS) balancing algorithms.

are inadequate for domains where fairness is critical. Fairness-aware preprocessing methods (Kamiran and Calders, 2012), on the other hand, target biases in sensitive attributes (e.g., gender, ethnicity, or age) by modifying data distributions, but operate on single-label datasets and do not incorporate multi-label class balancing. Recent efforts, such as SimFair (Liu et al., 2023), apply fairness considerations to multi-label scenarios by focusing on fairness constraints during evaluation rather than preprocessing. MLOPT-Ext bridges this gap by introducing a unified framework that simultaneously optimizes multi-label class balancing and demographic fairness. Table 5.1 summarizes the main features of MLOPT [P6] (Rieger et al., 2020b) compared to other competitive oversampling algorithms. In the table, *OS* stands for oversampling, *Bound* indicates whether the number of clones is controlled by dataset size, imbalance metric, or label-set frequency, and *Synthetic* refers to whether synthetic samples or labels are generated.

5.2 General Methodology

This section introduces the MLOPT algorithm for balancing multi-label datasets ([P6] (Rieger et al., 2020b)). MLOPT can be extended to optimize multiple objectives such as demographic metadata (MLOPT-Ext). MLOPT formulates multi-label class balancing as a nonlinear, non-convex constrained optimization problem with simple bound constraints. The formulation generalizes to multi-objective balancing (e.g., simultaneous balancing of label frequencies and demographic subgroups), referred to in this thesis as *MLOPT-Ext*.

The optimization is solved using a trust-region constrained method, which is a class of iterative local optimizers suitable for nonlinear objectives under box constraints. MLOPT oversamples label-sets subject to bounds on repetition, then counts are chosen that minimize a convexified imbalance measure. The following descrip-

Algorithm 1 MLOPT: Multi-Label Balancing Optimizer

- 1: **Input:** Dataset $\mathbf{D}$; oversampling bound $r > 1$; trade-off weight $\lambda \in [0, 1]$; distance metric $d(\cdot, \cdot)$
 - 2: **Output:** Balanced dataset $\mathbf{D}'$
 - 3: **Step 1: Create unique label-sets**
 - 4: $\mathbf{Y} \leftarrow$ unique label-set matrix extracted from $\mathbf{D}$ $\triangleright \mathbf{Y} \in \{0, 1\}^{m \times k}$
 - 5: $\mathbf{n}_o \leftarrow$ frequency of each label-set in $\mathbf{D}$ $\triangleright \mathbf{n}_o \in \mathbb{N}^m$
 - 6: **Step 2: Definition of oversampling bounds for each label-set**
 - 7: $\text{lower_bound} \leftarrow \mathbf{n}_o$
 - 8: $\text{upper_bound} \leftarrow r \cdot \mathbf{n}_o$
 - 9: **Step 3: Define objective**
 - 10: $\mathbf{n}_u :=$ oversampling plan (target counts per label-set)
 - 11: $\mathbf{z} \leftarrow \mathbf{Y}^\top \mathbf{n}_u$ $\triangleright$ resulting class counts
 - 12: $\boldsymbol{\mu}_z \leftarrow$ desired target class distribution $\triangleright$ e.g., uniform across classes
 - 13: $\text{variance_term} \leftarrow \text{Var}(\mathbf{n}_u \oslash \mathbf{n}_o)$
 - 14: $\text{distance_term} \leftarrow d(\mathbf{z}, \boldsymbol{\mu}_z)$
 - 15: $h(\mathbf{n}_u) \leftarrow \lambda \cdot \text{variance_term} + (1 - \lambda) \cdot \text{distance_term}$
 - 16: **Step 4: Constrained optimization problem**
$$\mathbf{n}_u^{opt} = \arg \min_{\mathbf{n}_u} h(\mathbf{n}_u) \quad \text{s.t.} \quad \mathbf{n}_o \leq \mathbf{n}_u \leq r \cdot \mathbf{n}_o$$
 - 17: **Step 5: Synthesis of balanced dataset**
 - 18: $\mathbf{D}' \leftarrow$ replicate samples of each label-set according to $\mathbf{n}_u^{opt}$ (with augmentation)
 - 19: **Return** $\mathbf{D}'$
-

tion presents a consolidated and notation-consistent formulation of the method, extending and refining the presentation in [P6].

5.2.1 MLOPT: Balancing Classes

The purpose of MLOPT is to construct a rebalanced training set such that underrepresented labels and underrepresented label combinations receive higher representation without disproportionately increasing other classes. Algorithm 1 provides an overview. The main algorithmic steps are detailed below.

1. **Create Unique Label-Sets:** Consider a multi-label classification problem with k classes (e.g., facial AUs). Each training sample is annotated with a binary indicator vector of length k , where a value of 1 indicates that the class is active for that sample.
Different samples may share identical label combinations. These identical

combinations are merged into *unique label-sets*. Let m denote the number of distinct label-sets observed in the dataset. The matrix

$$\mathbf{Y} \in \{0, 1\}^{m \times k}$$

is defined such that each row of $\mathbf{Y}$ corresponds to one unique label-set. The vector

$$\mathbf{n}_o \in \mathbb{N}^m$$

is defined such that $\mathbf{n}_o[i]$ is the number of samples in the original dataset that exhibit the i -th label-set.

Operating on unique label-sets rather than on individual samples has two main effects. First, it reduces computational cost, since the optimization allocates counts to m label-sets instead of to all samples. Second, it preserves higher-order co-occurrence structure. This is important for structured labels such as facial AUs, where certain combinations are not simple linear superpositions of individual AUs (see Section 2.2.2). As a result, the method can explicitly increase the frequency not only of rare classes, but also of rare *co-occurrences*.

2. **Definition of Oversampling Bounds:** The aim of the balancing procedure is to determine how often each label-set should appear in the rebalanced dataset. The target frequencies are defined in

$$\mathbf{n}_u \in \mathbb{N}^m,$$

where $\mathbf{n}_u[i]$ is the target number of samples that should carry label-set i after balancing.

To prevent uncontrolled replication, each label-set is assigned a lower and an upper bound. The lower bound enforces that no label-set is removed:

$$\mathbf{n}_o \leq \mathbf{n}_u.$$

The upper bound limits how aggressively rare label-sets may be oversampled. This upper bound is defined as a multiplicative factor r applied to the original counts:

$$\mathbf{n}_u \leq r \cdot \mathbf{n}_o, \quad r \in \mathbb{R}, \quad r > 1.$$

The factor r controls how many copies of a sample with a certain label-set are permitted. This constraint is important in practice because repeated duplication of a very small number of rare samples is known to increase overfitting in deep neural networks.

3. **Class Distribution Induced by Oversampling:** For any proposed vector $\mathbf{n}_u$, the resulting class-wise frequencies in the rebalanced dataset can be computed. The vector

$$\mathbf{z} = \mathbf{Y}^\top \mathbf{n}_u \in \mathbb{N}^k$$

is defined such that $\mathbf{z}[j]$ counts how often class j would appear after applying the oversampling frequencies in $\mathbf{n}_u$. Intuitively, $\mathbf{z}$ is the class histogram of the rebalanced dataset.

A target class distribution is also specified. This target is denoted by

$$\mu_z \in \mathbb{R}^k.$$

In the simplest case, μ_z can be chosen as the per-class mean count, i.e., all classes are encouraged toward the same frequency. More generally, μ_z may encode an application-specific desired distribution (for example, clinically meaningful priors or safety constraints).

4. **Objective Function:** The quality of the oversampling distribution $\mathbf{n}_u$ is measured by an objective with two terms: (1) a regularization term that penalizes highly uneven oversampling across label-sets, and (2) a distribution-matching term that encourages the resulting class frequencies to approach the desired target.

The first term is based on the element-wise ratio $\mathbf{n}_u \oslash \mathbf{n}_o$, where $\oslash$ denotes element-wise division. The quantity $\mathbf{n}_u[i]/\mathbf{n}_o[i]$ is the oversampling factor applied to label-set i . The variability of these factors across all label-sets is summarized by

$$\text{Var}(\mathbf{n}_u \oslash \mathbf{n}_o),$$

i.e., the variance of the oversampling factors. High variance indicates that some label-sets are being replicated far more aggressively than others.

The second term measures the discrepancy between the achieved class distribution $\mathbf{z}$ and the desired distribution μ_z . This discrepancy is written as

$$d(\mathbf{z}, \mu_z),$$

where $d(\cdot, \cdot)$ denotes a distance measure such as ℓ_1 , ℓ_2 , or Kullback–Leibler divergence. A smaller value of d indicates that the rebalanced dataset is closer to the target class proportions.

Both terms are combined into the objective

$$h(\mathbf{n}_u) = \lambda \text{Var}(\mathbf{n}_u \oslash \mathbf{n}_o) + (1 - \lambda) d(\mathbf{z}, \mu_z), \quad (5.1)$$

with a trade-off parameter

$$\lambda \in [0, 1].$$

Large values of λ emphasize an evenly spread oversampling. Small values of λ emphasize a higher movement toward the desired class distribution.

5. **Constrained Optimization Problem:** The task is to find oversampling frequencies $\mathbf{n}_u$ that minimize the objective in Equation 5.1, while respecting the admissible range for each label-set. The resulting optimization problem is

$$\mathbf{n}_u^{opt} = \arg \min_{\mathbf{n}_u} h(\mathbf{n}_u) \quad \text{subject to} \quad \mathbf{n}_o \leq \mathbf{n}_u \leq r \cdot \mathbf{n}_o. \quad (5.2)$$

The bounds $\mathbf{n}_o \leq \mathbf{n}_u \leq r \cdot \mathbf{n}_o$ form simple box constraints. The objective $h(\mathbf{n}_u)$ is nonlinear and, in general, non-convex because it contains both a variance term and a potentially non-convex distance term.

The optimization problem in Equation 5.2 is solved using a trust-region constrained algorithm (Conn et al., 2000). Trust-region methods generate iterative updates within a local region around the current candidate solution (the trust region). Within this region, a local approximation of the objective is considered reliable, and constraint handling (in this case, the box constraints on $\mathbf{n}_u$) is enforced directly.

Trust-region constrained optimization is appropriate in this setting for three reasons. First, the objective is nonlinear. Second, the feasible set is explicitly bounded. Third, stability of convergence to a feasible point is preferred over unconstrained descent that may leave the admissible range.

The output of this step is

$$\mathbf{n}_u^{opt} \in \mathbb{N}^m,$$

the optimized number of times each unique label-set should appear in the balanced dataset.

6. **Synthesis of Balanced Dataset:** A rebalanced dataset $\mathbf{D}'$ is then constructed by replicating samples according to $\mathbf{n}_u^{opt}$. For each label-set i , samples associated with this label-set are cloned until the target count $\mathbf{n}_u^{opt}[i]$ is reached. To avoid simple duplication, cloned samples are typically augmented using transformations that are valid for the given application. This step increases both the effective sample size of underrepresented label-sets and the variability within those label-sets.

The final dataset $\mathbf{D}'$ satisfies two goals: (1) rare classes and rare label-set combinations receive higher representation, and (2) oversampling is bounded

and regularized, limiting the risk of overfitting to a handful of rare training examples.

5.2.2 MLOPT-Ext: Balancing Multiple Objectives

MLOPT-Ext extends MLOPT from single-objective class balancing to multi-objective balancing. In addition to adjusting class frequencies, it also balances subject-level metadata such as gender, ethnicity, and age. This introduces domain knowledge directly into the resampling process and supports fairness-oriented datasets.

All factors to be balanced are combined into a single design matrix

$$\mathbf{Y} = [\mathbf{Y}_{AU} \mid \mathbf{Y}_G \mid \mathbf{Y}_E \mid \mathbf{Y}_A] \in \{0, 1\}^{m \times (k_{AU} + k_G + k_E + k_A)},$$

where each row corresponds to one unique label-set (Section 5.2.1), and the block matrices encode: AUs ($\mathbf{Y}_{AU}$), gender ($\mathbf{Y}_G$), ethnicity ($\mathbf{Y}_E$), and age ($\mathbf{Y}_A$). As in MLOPT, the optimizer selects target replication counts $\mathbf{n}_u \in \mathbb{N}^m$ for these label-sets with

$$\mathbf{n}_o \leq \mathbf{n}_u \leq r \cdot \mathbf{n}_o.$$

The resulting distributions are

$$\mathbf{z} = \mathbf{Y}_{AU}^\top \mathbf{n}_u, \quad \mathbf{g} = \mathbf{Y}_G^\top \mathbf{n}_u, \quad \mathbf{e} = \mathbf{Y}_E^\top \mathbf{n}_u, \quad \mathbf{a} = \mathbf{Y}_A^\top \mathbf{n}_u,$$

where $\mathbf{z}$ is the action-unit frequency vector in the balanced dataset, and $\mathbf{g}, \mathbf{e}, \mathbf{a}$ are the demographic frequency vectors (gender, ethnicity, age). For each of these, a desired target distribution is specified: μ_z, μ_g, μ_e , and μ_a .

MLOPT-Ext uses scalarization to combine all objectives into a single function:

$$\begin{aligned} h(\mathbf{n}_u) = & \lambda \text{Var}(\mathbf{n}_u \oslash \mathbf{n}_o) + (1 - \lambda - \eta) d(\mathbf{z}, \mu_z) \\ & + \eta \left(\frac{1}{k_G} d(\mathbf{g}, \mu_g) + \frac{1}{k_E} d(\mathbf{e}, \mu_e) + \frac{1}{k_A} d(\mathbf{a}, \mu_a) \right), \end{aligned} \quad (5.3)$$

where:

- λ and η are scalar weights controlling the trade-off between (1) oversampling stability, (2) class balance, and (3) demographic balance.
- k_G, k_E, k_A denote the number of categories for gender, ethnicity, and age, respectively. The factors $\frac{1}{k_G}, \frac{1}{k_E}$, and $\frac{1}{k_A}$ normalize the demographic terms.

Minimizing $h(\mathbf{n}_u)$ like MLOPT yields oversampling rates that simultaneously (1) reduce class imbalance, (2) align demographic proportions with specified targets, and (3) avoid replicating only a handful of rare cases. The result is a dataset that is

not only more balanced in terms of labels, but also more demographically equitable for downstream training.

5.3 Summary of Key Results

The proposed balancing framework was evaluated in two settings: (1) class imbalance reduction for multi-label data (MLOPT) and (2) joint balancing of class labels and demographic attributes (MLOPT-Ext). Additional quantitative results are provided in Appendix A.

Unless stated otherwise, learning setups, architectures, evaluation frameworks, metrics, and datasets follow Tables 2.5–3.1.

5.3.1 MLOPT: Superior to Competitive Algorithms on Benchmark Datasets

MLOPT was compared to established oversampling strategies (LPROS, MLROS) on standard multi-label benchmark datasets that are not specific to FER. The experiments are presented in Sections A.2 and A.3 in Appendix A.

The comparison controlled for differences in constraints: MLOPT was run with repetition bounds, and LPROS/MLROS with percentage-based oversampling limits, chosen to be equivalent in effect. In addition, all methods were evaluated under their preferred objectives.

Main findings:

- **Lower imbalance across all datasets.** MLOPT consistently achieved the lowest MeanIR, indicating flatter label-set distributions and reduced dominance of frequent combinations.
- **Robust under dataset artefacts.** MLOPT remained operational where baselines failed, such as cases where LPROS assumes a certain imbalance structure that does not hold, or MLROS cannot handle missing labels. This demonstrates higher practical robustness.
- **Objective choice matters.** L2-based objectives produced the strongest overall balancing, yielding more uniform label-set frequencies and improved MeanIR, even when MeanIR was not the explicit target.

5.3.2 MLOPT: Improved DNN Performance for In-the-wild AU Detection [P6]

The original MLOPT setting was applied to facial AU detection in-the-wild, as reported in [P6] (Rieger et al., 2020b).

Main findings:

- **Competitive leaderboard result.** A parameter-reduced ResNet-18 (ResNet18-112, Section 2.3.2) trained on MLOPT-balanced data achieved third place in the AU challenge of the Affective Behavior Analysis in the Wild (ABAW) challenge at the Face and Gesture (FG) conference (Kollias et al., 2020).
- **Better in-the-wild generalization.** On Aff-Wild2 validation and test sets, the balanced and augmented training data led to higher macro-F1 scores, improved detection of underrepresented AUs, and smaller validation–test performance gaps. The reduced gap indicates better generalization to unseen subjects and conditions.

5.3.3 MLOPT-Ext: Simultaneous Class and Demographic Balancing

MLOPT-Ext was evaluated on a facial affect dataset with associated demographic annotations (BP4D, Section 3.2). The results are presented in the Appendix Section A.3. The goal was to balance (1) AU label distributions and (2) subject demographics (gender, ethnicity, age) at the same time.

Main findings:

- **No fairness term ($\eta = 0$):** AU balance improves but demographic skew remains largely unchanged: the optimizer focuses only on AU coverage and ignores demographics when not penalized.
- **With fairness term ($\eta > 0$):** Demographic balance improves substantially without collapsing AU diversity. So, the dataset can be steered toward both goals.
- **L1 works best for demographics.** An L1 distance in the demographic component yields lower imbalance and spread across gender/ethnicity/age, while also improving AU MeanIR.
- **Best joint setting.** Moderate η with a balanced trade-off between AU and demographic terms achieves the strongest combined outcome.

These results demonstrate that fairness constraints can be directly enforced at the data-preparation stage, providing a strong foundation for training more equitable models.

5.4 Summary and Discussion

The research question **RQ2**: *How can multi-label datasets be balanced across multiple objectives in a way that remains suitable for training deep neural networks?* was addressed through the design of two algorithms: MLOPT and its extension MLOPT-Ext.

Label-set-level balancing for realism and control: MLOPT is tailored to multi-label data and operates at the level of unique label-sets (i.e., distinct co-occurrence patterns of labels). Instead of oversampling individual classes independently, it optimizes how often each label-set should appear in the final training set. This preserves realistic co-occurrences and prevents oversampling of a single rare combination. To remain suitable for deep learning, MLOPT enforces an upper bound on repetition via a multiplicative limit r . This bound restricts how often any label-set may be cloned and therefore limits overfitting due to aggressive augmentation of a few rare samples. As a result, MLOPT reduces imbalance without sacrificing variability, and supports better generalization in downstream neural networks.

Multi-objective extension for fairness-aware datasets: MLOPT-Ext generalizes this idea to multiple, potentially competing objectives. In addition to balancing task-relevant labels (e.g., AUs), MLOPT-Ext can simultaneously balance subject-level metadata such as gender, ethnicity, and age. This enables dataset construction that accounts not only for task performance but also for fairness. The scalarized objective function formalizes this by assigning explicit weights to task balance and demographic balance, making both targets optimizable and measurable.

Hyperparameters and practical workflow: The method depends on several hyperparameters. The weight λ controls how strongly the algorithm enforces the desired task label distribution. The weight η controls how strongly demographic balance is enforced. The bound r controls how aggressively rare label-sets may be replicated. Careful tuning of these parameters is important for a given dataset. However, this tuning is performed once at the dataset level. After a balanced version of the dataset has been generated, that dataset can be reused to train different deep neural network architectures without repeating the balancing procedure. This makes the approach practical for model development and for later deployment.

Applicability: The applicability of MLOPT(-Ext) is broad. It can be used with multi-label data beyond FER, and it remains effective even under severe label imbalance or incomplete annotations. At the same time, performance depends on the structure of the dataset: highly skewed label distributions, missing metadata, or extremely sparse co-occurrences can make the optimization problem harder and may require different hyperparameter settings.

Knowledge-Informed Model

Training: CorrLoss and FMC-Net [P4, P7]

This chapter addresses **RQ3**: *How can relationships between classes and tasks be enforced during training?*

The scientific contributions, visualizations, and results discussed in this chapter are derived from my publications [P4] (Rieger et al., 2022) and [P7] (Rieger et al., 2025).

This chapter introduces a training-time constraint. The central idea is to augment standard supervised losses with a regularizer that penalizes deviations from target correlation structure. Two sources of domain knowledge are encoded: (1) AU–AU co-activation patterns and (2) AU–FE associations reflecting the hierarchical relation between AUs and categorical FEs. Target correlations can be derived from empirical statistics or from established FACS-based priors. Within this unified view, CorrLoss [P4] is the single-task instantiation that constrains AU–AU structure during AU detection. FMC-Net [P7] generalizes the constraint to multi-task learning by enforcing AU↔FE consistency (and optionally AU–AU) during training. Inference-time refinements in FMC-Net with rule-based fuzzy logic and an risk-aware ensemble are presented separately in Chapter 7. This chapter begins with a positioning of my contributions, followed by a brief summary of the unified algorithm and a summary of the empirical results demonstrating improved robustness, reduced task interference and better generalization in both single-task and multi-task settings.

This chapter incorporates domain knowledge in the form of AU co-occurrences (Section 2.2.2) and the concept of hierarchical task relationships (Section 2.2.3).

6.1 Motivation and Related Work

While data-centric measures (Chapter 5) address imbalance and bias, the model training objective is a direct lever for improving robustness and adaptability in FER.

The FACS (Ekman and Friesen, 1978) defines AUs as muscular primitives underlying basic emotions (Friesen, Ekman, et al., 1983) and finer-grained expressions (Kunz and Lautenbacher, 2014; Hassan et al., 2019; Shao et al., 2024). Two regularities are especially useful during training: (1) AU–AU co-activation patterns (physiology-informed co-occurrence, see Section 2.2.2) and (2) AU–FE associations linking AUs to categorical facial expressions (hierarchical task relations, see Section 2.2.3). These target correlations provide principled signals to reduce brittleness in purely data-driven models (Cui et al., 2020; Yang et al., 2021; Dash et al., 2022; Li and Shan, 2023). Both correlation types are expected to generalize beyond a single dataset.

Human-centered and interactive ML argue for inspectable, steerable models (Shneiderman, 2022; Holzinger, 2016). In FER, existing systems often adapt interfaces or lexicons (Alipour et al., 2023; Gamage et al., 2024) rather than the models internal variables. Binding domain knowledge to AUs such as in the FMC-Net [P7] (Rieger et al., 2025) can shape learned representations and thereby support both robustness and controlled, interpretable adaptation.

To shape the models internal variables, structured knowledge has been injected via hypothesis spaces (knowledge graphs, Bayesian relations) for AU and AU→FE dependencies (Li et al., 2019; Song et al., 2021a; Cui et al., 2020), and via patch-level or multi-label learning to capture correlations (Corneanu et al., 2018; Zhao et al., 2015). Unlike existing methods, CorrLoss [P4] (Rieger et al., 2022) enforces target correlations directly via the loss.

Correlation coefficients (PCC/CCC) are common for evaluation or auxiliary design (Ringeval et al., 2015; Atmaja and Akagi, 2021; Song et al., 2021b). CorrLoss [P4] (Rieger et al., 2022) advances this by introducing a correlation-discrepancy term as a differentiable training regularizer that enforces AU–AU structure, capturing probabilistic relationships and improving cross-dataset robustness.

Multi-task learning (MTL) uses a shared representations for AU and FE (Zhang et al., 2022; Zhang et al., 2023b), but can suffer negative transfer (performance drop) without explicit alignment. Addressing this, FMC-Net [P7] (Rieger et al., 2025) extends CorrLoss by enforcing inter-task $AU \leftrightarrow FE$ consistency, providing a direct, tunable constraint for joint training.

For a detailed discussion of prior work, see the related work sections of [P4] and [P7].

6.2 General Methodology

This section introduces the general methodology for correlation-informed training as a unified approach derived from [P4] (Rieger et al., 2022) and [P7] (Rieger et al.,

2025).

This work introduces a general methodology for injecting domain structure into training by penalizing discrepancies between predicted and target correlation patterns. Two cases are considered:

1. AU–AU structure for single-task AU detection (CorrLoss, [P4]), where Figure 6.1 gives an intuition for the AU–AU constraint on happy FEs.
2. AU↔FE structure for joint AU and FER (FMC-Net, [P7]), where Figure 6.2 illustrates this constraint-guided multi-task training in FMC-Net. Inference-time components (fuzzy logic and the ensemble) of the FMC-Net are described in Chapter 7.

6.2.1 Correlation Constraints

The regularization term is introduced as part of the loss function during the training. For a batch of size B with A (AU) and F (FE) classes, the network outputs $\hat{Y}^{(A)} \in \mathbb{R}^{B \times A}$ (sigmoid) and $\hat{Y}^{(F)} \in \mathbb{R}^{B \times F}$ (softmax). The supervised losses are the binary cross-entropy L_A (AUs) and the categorical cross-entropy L_F (FEs). They are averaged over the training batch.

Batchwise correlation matrices $\hat{C}_{AF}$ or $\hat{C}_{AA}$ are computed column-wise across samples using correlation coefficients (see Section 2.4.3):

$$\hat{C}_{AF} = \text{corr}(\hat{Y}^{(A)}, \hat{Y}^{(F)}) \in \mathbb{R}^{A \times F}, \quad \hat{C}_{AA} = \text{corr}(\hat{Y}^{(A)}, \hat{Y}^{(A)}) \in \mathbb{R}^{A \times A}. \quad (6.1)$$

Given pre-defined target matrices C_{AF}^* (AU–FE) and C_{AA}^* (AU–AU), p -norm discrepancies are defined over unique AU–AU pairs and all AU–FE pairs

$$R_{\text{task}} = \left(\frac{1}{AF} \sum_{a=1}^A \sum_{f=1}^F |\hat{C}_{AF}[a, f] - C_{AF}^*[a, f]|^p \right)^{1/p}. \quad (6.2)$$

$$R_A = \left(\frac{2}{A(A-1)} \sum_{1 \leq a < b \leq A} |\hat{C}_{AA}[a, b] - C_{AA}^*[a, b]|^p \right)^{1/p}. \quad (6.3)$$

with $p \in \{1, 2\}$. Here, R_{task} denotes the regularization term for AU–FE and R_A the term for AU–AU. The target correlation may be derived from training data, transferred from an auxiliary corpus, or specified from FACS-based priors.

6.2.2 Unified Objective and Single-Task Case

The task supervision and regularization term are optimized via

$$\tilde{L} = (1 - \alpha)(L_F + L_A) + \alpha(R_{\text{task}} + \gamma R_A), \quad \alpha \geq 0, \gamma \in \{0, 1\}. \quad (6.4)$$

Here, the hyperparameter α controls the overall strength of the structural regularization, while the binary flag γ allows the model to enforce only inter-task consistency ($\gamma = 0$) or both inter- and intra-task structure simultaneously ($\gamma = 1$).

Single-task case (CorrLoss). For AU detection alone ($L_F = 0$ and no AU-FE term),

$$\tilde{L}_{\text{AU}} = L_A + \alpha R_A, \quad (6.5)$$

which encourages physiologically plausible AU co-activations.

Interpretation. Minimizing these penalties reduces the discrepancy between the correlation structure between the predictions and the targets, aligning learned representations with facial muscle behavior (AU–AU) and the hierarchical AU→FE mapping in multi-task training.

6.3 Summary of Key Results

The empirical results from publications [P4] and [P7] validate the effectiveness of the correlation-informed regularization framework in both single-task and multi-task settings. Across settings, enforcing target correlation structure reduced variance, mitigated overfitting, and improved generalization.

Unless stated otherwise, learning setups, architectures, evaluation frameworks, metrics, and datasets follow Tables 2.5–3.1.

6.3.1 CorrLoss: AU–AU Constraint Improves Performance and Generalizability in AU Detection [P4]

CorrLoss augments AU detection with an AU–AU correlation discrepancy term computed batchwise from predictions in this experiment. Models were trained either on lab or in-the-wild datasets and evaluated both within-dataset and cross-dataset to assess transfer.

Main findings:

- **Within-dataset:** modest but consistent mean improvements in AU metrics, accompanied by notable variance reduction across folds; training curves indicated reduced overfitting.
- **Cross-dataset:** higher mean performance and lower variance, indicating improved transfer when AU co-activation structure is enforced during training.
- **Complementarity with balancing:** combining CorrLoss with MLOPT ([P6]) further improved results on two of three training datasets, suggesting that distributional rebalancing and structural regularization result in different error modes.

The results show that penalizing deviations from AU co-activation targets promotes representations that respect stable physiological co-occurrences, showing more robust behavior under dataset shift while preserving (and sometimes improving) in-distribution accuracy.

6.3.2 FMC-Net: $AU \leftrightarrow FE$ Constraint Improves Performance for Joint AU–FE Prediction [P7]

FMC-Net extends correlation-informed training to multi-task learning by enforcing $AU \leftrightarrow FE$ consistency (with an optional AU–AU term). Across the evaluated settings, the constraints improved the combined performance $(AU_{F1} + FE_{F1})/2$ and reduced variability across folds compared to unconstrained baselines.

Main findings:

- **Mean and stability.** Variants using the Pearson correlation with $L2$ (PCC+ $L2$) most often achieved the highest *mean* combined scores, whereas Concordance-based variants with $L2$ (CCC+ $L2$) tended to yield the lowest standard deviations across folds, i.e., more stable performance. The inter-task constraint (R_{task}) alone delivered consistent gains, while adding the intra-task AU term (R_{both}) further improved stability.
- **In-the-wild robustness.** On in-the-wild evaluation, R_{both} (PCC+ $L2$) produced substantial gains for FE with low dispersion, indicating that jointly aligning AU–FE and AU–AU structure stabilizes learning under real-world variability.
- **Task asymmetry.** Improvements were larger and more consistent for FE than for AU (AU improved in most but not all settings), consistent with the hypothesis that anatomically grounded AUs provide a useful inductive bias for FE recognition.

When accuracy is prioritized, PCC+ $L2$ with R_{task} (or R_{both}) is preferred, whereas when reliability across subjects is critical, CCC+ $L2$ variants provide lower dispersion with competitive means. Overall, explicit AU $\leftrightarrow$ FE consistency reduces negative transfer in multi-task FER while improving both average performance and stability.

6.4 Summary and Discussion

This work addresses **RQ3**. *How can relationships between classes and tasks be enforced during training?* by introducing correlation-informed regularization CorrLoss for AU–AU structure in single-task AU detection and a task-constraint for AU $\leftrightarrow$ FE consistency within FMC-Net for multi-task learning.

Effectiveness and stability: Embedding correlation targets into the loss improved robustness and generalization beyond cross-entropy alone. For AU detection, CorrLoss reduced fold-to-fold variance and improved cross-dataset transfer. Furthermore, it indicated mitigation of overfitting via structure consistent with physiological co-activations. In multi-task settings, the task-constraint aligned AU and FE heads, reduced negative transfer, and increased the combined score $(AU_{F1} + FE_{F1})/2$. Mean–variance analyzes showed a useful trade-off: PCC+ $L2$ variants most often yielded the highest means, while CCC+ $L2$ variants tended to exhibit lower standard deviations across subject-dependent folds (i.e., greater stability). Adding AU–AU to AU $\leftrightarrow$ FE (R_{both}) further stabilized performance in settings with higher label noise or distribution shift.

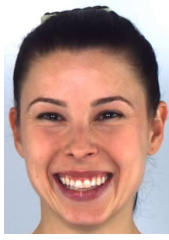
Interpretation: Penalizing discrepancies between predicted and target correlations constitutes an explicit inductive bias that grounds domain knowledge in internal variables (AUs, FEs). The constraint encourages features that respect non-additive AU relations and the hierarchical AU $\rightarrow$ FE mapping, thereby improving transfer without relying solely on data rebalancing. The larger and more consistent gains for FE than AU suggest that anatomically grounded AUs provide informative guidance for FE recognition in multi-task training.

Design choices and tuning: Targets can be derived from empirical training statistics, transferred from representative corpora, or specified via FACS-based priors. The weighting α governs the strength of knowledge injection; appropriate selection on validation data is critical for managing the bias–variance trade-off. In practice, PCC+ $L2$ is preferable when peak mean performance is prioritized; CCC+ $L2$ is advantageous when reliability across subjects (low dispersion) is essential. The AU–AU term is most beneficial when AU labels are noisy or imbalanced, complementing the inter-task constraint.

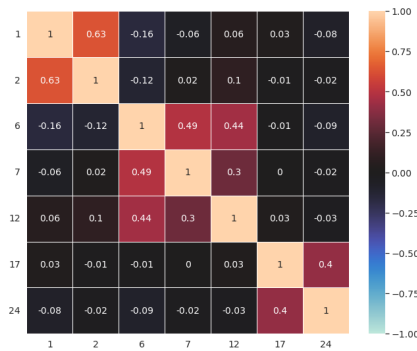
Scope, limitations, and adaptability: Batchwise correlation estimates provide a

lightweight, differentiable mechanism but remain sensitive to small batch statistics and to misspecified targets; consequently, α and target provenance require scrutiny. The explicit form of the constraints, however, enables controlled adaptation and transparency: updating the target matrices offers a symbol-grounded way to tailor behavior to dataset shifts or application norms without architectural changes.

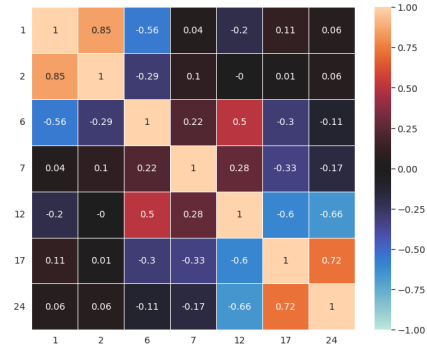
Broader applicability: Correlation-informed constraints generalize to tasks with probabilistic or hierarchical dependencies. Within affect recognition, extensions include constraints for appraisal dimensions (e.g., valence/novelty/control), pain or depression indicators expressible via AU constellations, or FE–FE relations in compound emotions. Beyond FER, analogous formulations apply to hierarchical image classification, multi-label industrial inspection, or multimodal diagnostics where symptom/task relations are known. Overall, correlation-informed regularization offers a compact, adaptable mechanism to inject domain structure, improving average performance, stability, and interpretability in settings with limited or biased data.



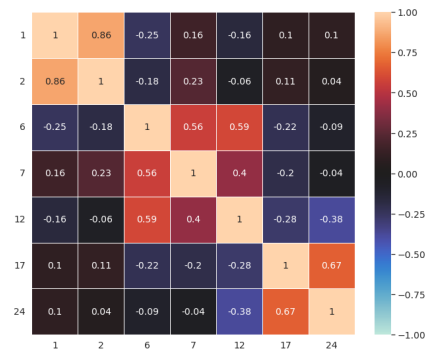
(a): Happy face example from (Seuss et al., 2019).



(b): Target AU correlations for happiness. There is eye brow movement (AU 1-2), cheek raising and lid tightening (AU 6-7), and mouth movements that extend also to the cheek and chin (AU 12-24).



(c): Predicted correlations when trained with **binary cross-entropy (BCE)** only.



(d): Predicted correlations with **BCE + CorrLoss**. This matrix is more similar to the target correlation compared to when only trained with BCE.

Fig. 6.1.: CorrLoss (d) brings predicted AU–AU correlations closer to target correlations (b) (adapted from [P4] (Rieger et al., 2022)). This is calculated for the FE happy on the BP4D dataset.

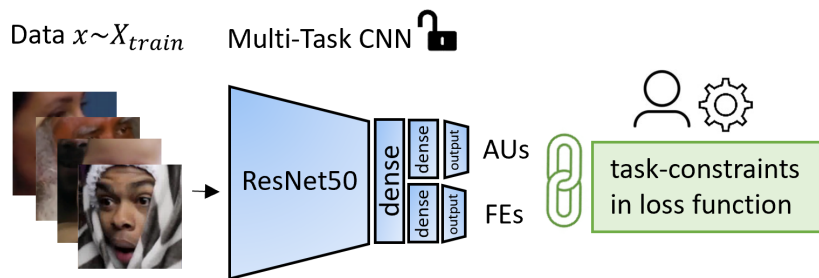


Fig. 6.2.: Constraint-guided training in FMC-Net for emotion recognition: a multi-task CNN predicts AUs and FEs while a differentiable term encodes AU–FE relationships to regularize shared representations. Adapted from [P7] (Rieger et al., 2025).

Rule- and Model-Based Ensemble for Transparent Inference: FMC-Net [P7]

This chapter addresses *RQ4: How can an expert-informed ensemble be designed for facial expression recognition?*

The scientific contributions, visualizations, and results discussed in this chapter are derived from my publication [P7] (Rieger et al., 2025).

Building on the training-time constraints of Chapter 6, the facial movement constraint network (FMC-Net) is extended at inference with a second, complementary layer of human guidance: a rule–model ensemble. The ensemble combines (1) a model-based FE prediction from the multi-task CNN and (2) a fuzzy rule-based FE prediction obtained by mapping AU probabilities to emotions via editable AU→FE rules. This serves three purposes. First, it improves performance by using complementary error profiles of the symbolic and neural pathways. Second, it enables disagreement-based risk quantification: differences between the two predictors yield an interpretable signal suitable for selective prediction and post-deployment monitoring. Third, the rule pathway provides user-editable logic, supporting transparent, domain-informed adaptation without retraining.

This chapter incorporates domain knowledge in the form of hierarchical task relationships expressed as logic rules for mapping AUs to emotions (see Table 2.4 in Section 2.2.3).

7.1 Motivation and Related Work

Mapping AUs to higher-level FEs has long used rule-based formulations, from statistical relations and templates (Velusamy et al., 2011) to Boolean rules derived from psychology (Valstar and Pantic, 2006; Ruiz et al., 2015; Seuss et al., 2023). Such approaches provide structure but are often fixed or opaque to end users. To interface symbolic knowledge with continuous DNN outputs, fuzzy logic offers graded

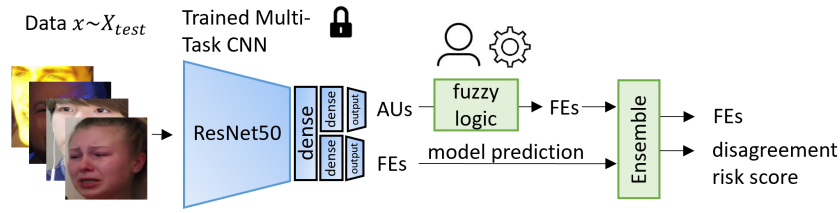


Fig. 7.1.: FMC-Net inference with a rule–model ensemble. AU probabilities from the AU head are mapped to emotions by a fuzzy, user-editable AU-to-FE rule layer, while a parallel CNN head provides model-based emotion predictions. The ensemble combines both sources to produce the final decision and an interpretable risk score based on their disagreement. Adapted from [P7] (Rieger et al., 2025).

memberships and t-norm operators (Hájek et al., 1996), enabling differentiable or modular integration with deep predictors (Badreddine et al., 2022). Recent applications include concept verification in vision (Schwalbe et al., 2022) and refinement of pseudo-labels (Liang et al., 2023). FMC-Net builds on this idea to provide a transparent, user-editable AU→FE pathway grounded in FACS-based rules.

Ensembles are widely used to improve robustness by aggregating complementary predictors (Lakshminarayanan et al., 2017; Gawlikowski et al., 2023), including in FER (Zhang et al., 2024a; Pons and Masip, 2018). Beyond averaging, the disagreement between heterogeneous predictors can serve as an actionable risk signal: when a symbolic, rule-guided pathway and a neural pathway diverge, the conflict localizes potential failure modes to specific AU rule activations, supporting selective prediction and post-deployment monitoring.

FMC-Nets ensemble ([P7] (Rieger et al., 2025)) instantiates these principles by combining (1) a model-based FE prediction from a multi-task CNN and (2) a fuzzy rule-based FE prediction driven by AU probabilities and editable AU→FE rules. The rule pathway confers transparency and editability; the neural pathway contributes data-driven generalization. Their agreement provides confidence calibration, and their disagreement yields an interpretable risk cue. This design addresses limits of purely data-driven systems—lack of case-wise rationale and limited post-hoc controllability, while avoiding brittleness of purely rule-based systems through complementary model evidence.

7.2 General Methodology

This section outlines FMC-Net’s inference-time components (Figure 7.1): a fuzzy AU→FE mapper and a rule–model ensemble with a disagreement-based risk signal.

7.2.1 Fuzzy AU→FE Mapping (Rule-Based Inference)

Let $\mathbf{a} \in [0, 1]^A$ denote AU probabilities from the AU head. AU→FE relations are encoded as differentiable fuzzy formulas. Product fuzzy logic (Hájek et al., 1996) is used with

$$\text{fuzzy-AND } (x, y) = xy, \quad \text{fuzzy-OR } (x, y) = x + y - xy,$$

applied to AU literals. For each emotion $f \in \{1, \dots, F\}$ a raw rule score $s_f(\mathbf{a}) \in [0, 1]$ is obtained by composing these operators over the rule's AU terms. Raw scores are converted to a probability vector by ℓ_1 -normalization:

$$p_{\text{rule}}(f) = \frac{s_f}{\sum_{g=1}^F s_g}, \quad p_{\text{rule}} \in \Delta^{F-1}.$$

This yields a transparent, user-editable FE predictor. Product fuzzy logic is compatible with continuous model outputs such as from AUs (Badreddine et al., 2022).

7.2.2 Rule–Model Ensemble and Disagreement-Based Risk

Let $p_{\text{model}} \in \Delta^{F-1}$ be the FE distribution from the CNN head (softmax). The ensemble forms a convex mixture

$$p_{\text{ens}} = (1 - \eta) p_{\text{model}} + \eta p_{\text{rule}}, \quad \eta \in [0, 1],$$

with default $\eta = \frac{1}{2}$. The predicted class is $\arg \max_f p_{\text{ens}}(f)$.

Disagreement-based risk signal. This is operationalized as the divergence between the neural and rule predictions:

$$d(f) = |p_{\text{model}}(f) - p_{\text{rule}}(f)|, \quad s_{\text{inst}} = \frac{1}{2} \sum_{f=1}^F d(f) \in [0, 1].$$

Here, s_{inst} is the total variation distance between p_{model} and p_{rule} , where higher values indicate stronger rule–model disagreement. Selective prediction could accept a decision when $s_{\text{inst}} \leq \tau$ (determined threshold). For analysis, the per-class terms $d(f)$ localize conflicts to specific AU rules without introducing an additional scalar.

7.3 Summary of Key Results: Increasing Robustness and Interpretability for Emotion Recognition [P7]

The ensemble layer of FMC-Net, combining a fuzzy AU-to-FE rule head with a neural FE head, improves robustness and transparency across within- and cross-dataset settings. The fused predictor yields more stable performance than either component alone, while the rule–model disagreement provides an interpretable, deployable risk signal for selective prediction and post-deployment monitoring.

Unless stated otherwise, learning setups, architectures, evaluation frameworks, metrics, and datasets follow Tables 2.5–3.1.

Main findings:

- **Fuzzy rules over AU probabilities outperform Boolean rules.** Using Product fuzzy logic on continuous AU outputs yields higher and more stable FE performance than Boolean formulations, while remaining transparent and editable.
- **Effects are class-dependent** Stricter rules with multiple, complex AU combinations (e.g., for fear) are more prone to failure than simpler rules, suggesting they may require expert refinement.
- **Training-time task constraints strengthen the rule pathway.** When the AU head is trained with the correlation-based task constraint, rule-based FE predictions improve, indicating that better-structured AU outputs transfer to a more reliable AU→FE mapping.
- **Rule–model ensemble improves robustness.** Combining rule and model predictions provides balanced within-dataset performance and competitive cross-dataset results.
- **Disagreement-based risk is actionable.** The distributional disagreement between rule and model correlates with accuracy. This can support selective prediction and post-deployment monitoring, while remaining interpretable at the AU-rule level.
- **Efficiency and applicability.** The multi-task setup solves AU and FE jointly with modest overhead; rules can be edited without retraining, enabling practical adaptation and case-wise inspection.

7.4 Summary and Discussion

This chapter addressed **RQ4**: *How can an expert-informed ensemble be designed for facial expression recognition?* by integrating a fuzzy, rule-based AU→FE pathway with a neural FE head inside FMC-Net. The resulting rule–model ensemble improves robustness and transparency across within- and cross-dataset settings, while exposing an interpretable, deployable risk signal based on rule–model disagreement.

Robustness and transparency: By grounding inference in editable AU→FE rules, the ensemble preserves a stable, human-readable structure that complements the data-driven head. The fuzzy rule pathway uses continuous AU probabilities rather than binarized detections, thereby capturing graded evidence and yielding stronger and more stable FE predictions than with Boolean logic. The ensemble further balances complementary error profiles of the two heads and maintains performance under distribution shift, while providing case-wise rationales via inspected rule activations.

Actionable reliability signal: Disagreement between the rule and model distributions tracks accuracy in held-out and cross-dataset evaluation and can serve as a practical accept/reject criterion for selective prediction and post-deployment monitoring. Unlike confidence scores, this signal is semantically attributable: conflicts can be traced to specific AU rules and updated without retraining, aligning with human-guided adaptation goals.

Interaction with training-time constraints: The ensemble benefits from correlation-informed multi-task training (Chapter 6). Structuring the AU head with task constraints produces AU probabilities that transfer into more reliable rule-based FE outputs, reducing the gap between single- and multi-task performance and stabilizing the ensemble.

Limitations and failure modes: Error propagation from AU detection remains a risk, particularly for strict or conjunctive rules. Class-dependent effects were observed; for example, fear rules relying on multiple simultaneous AUs proved brittle and may require softening or context-aware refinement. These findings motivate small-scale rule audits focused on rare or compositionally complex expressions and, where appropriate, domain-specific variants.

Efficiency and deployment: The approach adds minimal computational overhead: a single multi-task model supplies both AU and FE, and the rule pathway is lightweight. Rules are editable post-deployment, enabling rapid adaptation to cohort- or application-specific requirements without model retraining.

Scope for generalization: Because the rule pathway consumes AU probabilities and expresses hierarchical relations, the method extends naturally to additional FER targets (e.g., appraisal dimensions (Seuss et al., 2023), compound emotions (Li

et al., 2023; Siebers et al., 2016)) and related affective states (e.g., pain (Kunz and Lautenbacher, 2014), depression (Gavrilescu and Vizireanu, 2019)), as well as to other domains with known concept hierarchies. This positions the ensemble as a reusable pattern for combining symbolic priors with neural inference in settings that demand robustness, interpretability, and controllable behavior.

Model Evaluation: Relevance-Guided Local Explanations (RelEx) [P1, P2, P5]

This chapter addresses **RQ5**: *How can the alignment between model predictions and physiologically plausible Action Unit locations be quantified for model validation?*

The scientific contributions, visualizations, and results discussed in this chapter are derived from my publications [P5] (**Rieger** et al., 2020a), [P2] (Finzel et al., 2023), and [P1] (**Rieger**, 2024).

Relevance-Guided Local Explanations (RelEx) is introduced as a post-hoc, domain-specific evaluation framework that measures whether the evidence used by a model for an AU decision is concentrated in anatomically plausible facial regions. The method operates on saliency maps produced by standard gradient- or relevance-based explainers and aggregates evidence within AU-specific regions derived from FACS-based localization priors. The resulting alignment scores complement conventional metrics by quantifying spatial validity of the underlying rationale.

Publications [P5] (**Rieger** et al., 2020a) and [P2] (Finzel et al., 2023) establish the RelEx framework and its application to AU classification models. Publication [P1] (**Rieger**, 2024) extends this by using RelEx to evaluate models trained with correlation-informed constraints (Chapter 6), thereby evaluating whether gains in generalization coincide with more anatomically aligned explanations. Overall, RelEx provides a tool for model selection, diagnosis, and reporting when physiological plausibility is a requirement.

This chapter incorporates domain knowledge in the form of localization of facial AUs (Section 2.2.1).

8.1 Motivation and Related Work

Deep neural networks in affect recognition remain opaque, which is particularly problematic in sensitive domains such as healthcare and psychological assessment where stakeholders require transparent, auditable behavior and justifiable predictions (Schmid, 2024; Bruckert et al., 2020). Explainable AI (XAI) has emerged in response to these demands, offering a variety of techniques that make model behavior more inspectable, ranging from global abstractions such as rule summaries and prototype-based characterizations of class structure (Rabold et al., 2020; Rabold et al., 2022) to local methods that account for individual decisions, including counterfactual explanations and other instance-level analyzes (Guidotti, 2024). Within computer vision, saliency-based techniques such as Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) and Grad-CAM (Selvaraju et al., 2017) highlight image regions that support a prediction. See background Section 2.4.2 for short method explanation. Such methods have revealed failure modes where models appear correct for the wrong reasons, e.g., relying on background watermarks rather than object features (Lapuschkin et al., 2019). These methods are therefore widely used for qualitative debugging, including in AU detection where they help visualize the facial areas driving a models decision (Zhao et al., 2016).

The primary limitation that motivates the present contribution is that saliency maps are effective for visually validating and debugging selected images; however, manually evaluating large numbers of instances becomes impractical. This highlights the need for systematic and quantitative evaluation approaches. However, existing quantitative protocols largely emphasise task-agnostic faithfulness of models. For example, there are approaches for perturbation-based Area Over the Perturbation Curve (Samek et al., 2017) or Faithfulness Correlation (Alvarez Melis and Jaakkola, 2018), which do not integrate domain knowledge.

To address this gap, Relevance-Guided Local Explanations (RelEx) (Rieger et al., 2020a; Finzel et al., 2023; Rieger, 2024) is introduced as a post-hoc, model- and explainer-agnostic audit that accepts any trained AU detector together with any saliency method that yields per-image relevance (e.g., LRP, Grad-CAM) and converts these heatmaps into AU-specific plausibility scores, thereby transforming qualitative heatmaps into quantitative, class-wise alignment measures suitable for systematic comparison across architectures, datasets, and training regimes.

A second limitation of saliency techniques is about the fidelity and robustness of these explainers, i.e. how accurately they reflect the model's true reasoning and how stable they are under slight input perturbations (Ali et al., 2023; Seth and Sankarapu, 2025). This does not preclude the use of RelEx, because RelEx remains agnostic to the particular explanation algorithm, such that fidelity diagnostics can

and should be applied to the explainer itself, after which RelEx provides a complementary, task-aware metric that scales beyond manual inspection and anchors model evaluation in domain anatomy.

8.2 General Methodology

This section outlines Relevance-Guided Local Explanations (RelEx), a post-hoc, model- and explainer-agnostic procedure to quantify whether class-specific saliency concentrates in physiologically plausible facial regions (Fig. 8.1). RelEx turns qualitative heatmaps into AU-wise plausibility scores that can be averaged across images, classes, and splits for systematic comparison of models, training regimes, and datasets. The RelEx method involves the following steps:

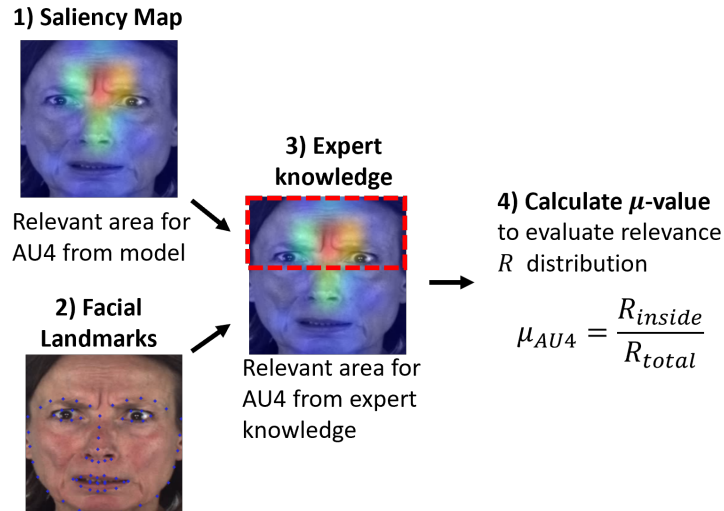


Fig. 8.1.: RelEx pipeline. (1) Class-specific saliency map (here: AU4 Brow Lowerer) from any local explainer. (2) Facial landmarks define facial subregions. (3) Domain knowledge (AU-localisation) defines AU-specific regions of interest (ROIs). (4) Quantification: the μ -score measures the fraction of relevance inside the ROI. Adapted from [Rieger \(2024\)](#).

1) Saliency extraction (class-conditional). Given a trained AU detector and a target AU, image saliency maps are computed with any local explanation method that produces relevance over pixels (e.g., LRP (Bach et al., 2015), Grad-CAM (Selvaraju et al., 2017)).

2) Landmark-based facial partitioning. Facial landmarks provide a subject-specific partition of the face into stable subregions. In the publications, SHORE™ (68 points) is used to obtain eyes, eyebrows, nose, mouth, and surrounding areas (Section 3.3).

3) AU-specific regions of interest (ROIs). Domain knowledge about AU localisation (Section 2.2.1) is instantiated as ROIs per AU. Two parametrisations are considered: (1) coarse bounding boxes (Rieger et al., 2020a) and (2) fine-grained polygons conforming to facial anatomy (Finzel et al., 2023). ROIs are applied per image using the detected landmarks, yielding subject- and pose-adapted masks.

4) Plausibility quantification (μ -score). Let R_{inside} denote the sum of saliency within the AU ROI and R_{total} the total saliency over the face crop. The AU-wise plausibility score is

$$\mu = \frac{R_{\text{inside}}}{R_{\text{total}}}.$$

A high μ -score indicates that the network’s decision is based on the expected region of interest and aligns with the expert knowledge. A low μ -score means that the network might be relying on irrelevant or spurious features outside the expected region, potentially leading to unreliable or biased decisions. This approach is particularly important in high-stakes domains such as healthcare, where it is important to validate the decision making of AI models to build trust and ensure system robustness.

8.3 Summary of Key Results

This section summarizes how RelEx quantifies local explanations to audit AU models and training regimes. Across publications, RelEx turns qualitative heatmaps into AU-specific plausibility scores that reveal when models rely on spurious context rather than physiologically expected regions.

Unless stated otherwise, learning setups, architectures, evaluation frameworks, metrics, and datasets follow Tables 2.5–3.1.

8.3.1 Unraveling Decisions of Trained CNNs for Action Unit Detection [P2, P5]

RelEx was applied to saliency maps generated with LRP to assess whether relevance concentrates in AU-specific facial regions.

Main findings:

- **Right-for-the-wrong-reasons effect.** Even for correctly predicted AUs, less than 50% of the positive relevance typically falls inside physiologically expected AU regions, indicating reliance on contextual or spurious cues despite high model performance. This underscores that, although these models performed well, their underlying decision-making processes did not consistently align with established domain knowledge.
- **Granularity matters.** Moving from coarse bounding boxes ([P5]) to fine-grained, polygonal AU regions ([P2]) yields more discriminative plausibility estimates.
- **Knowledge quality influences audits.** Results remain robust in trend but depend on the precision of expert-defined AU regions, underscoring the value of careful, anatomy-based priors.

8.3.2 Effect of Co-occurrence Regularization on Trained Features for Action Unit Detection [P1]

RelEx is used as an analysis tool to quantify how correlation-informed regularization (CorrLoss) changes on what internal features the trained AU models rely on. Rather than reporting only task performance, RelEx converts Grad-CAM saliency into AU-specific plausibility measures, allowing to compare models by how well their relevance aligns with physiologically expected AU regions.

Main findings:

- **Higher model performance with more distributed features.** $\text{CorrLoss}(L_2)$ achieves the highest average F1 on both Actor Study and BP4D while yielding the lowest average RelEx μ -values across the tested configurations. Since μ measures concentration inside a AU ROI, a lower μ here reflects a spread of relevance into secondary, physiologically implicated areas (co-activated AUs), consistent with the objective of enforcing AU correlation structure.
- **Regularizer choice matters.** Both CorrLoss variants alter model features in comparison to no regularization, but the L_2 distance yields the strongest performance gains together with a shift toward broader, multi-region saliency, indicative of using AU co-occurrence patterns.
- **Class-wise variability.** RelEx reveals systematic AU differences (e.g., lower μ for AU6/AU7), highlighting that correlation-informed training interacts with AU-specific anatomy and should be assessed per class.

8.4 Summary and Discussion

This chapter addressed **RQ5**: *How can the alignment between model predictions and physiologically plausible Action Unit locations be quantified for model validation?* by introducing and applying Relevance-Guided Local Explanations (RelEx) (Rieger et al., 2020a; Finzel et al., 2023). RelEx is a post-hoc, model- and explainer-agnostic audit that couples saliency maps with domain knowledge on AU localization to yield quantitative, AU-specific plausibility scores. Beyond auditing baseline models, RelEx was also used to analyze how correlation-informed regularization (CorrLoss) reshapes features in trained AU models (Rieger, 2024), thereby connecting explainability with training-time inductive biases.

What RelEx unravels about accuracy–plausibility gap and effects of CorrLoss on model features: Across experiments, RelEx reveals a systematic gap between standard performance metrics and physiologically plausible evidence use: models with strong performance can still focus substantially outside expected AU regions, i.e., they can be right for the wrong reasons (Rieger et al., 2020a). RelEx makes this gap measurable via the μ -score rather than relying on visual inspection alone. When applied to CorrLoss, RelEx further shows that the L_2 variant simultaneously improves predictive performance and distributes relevance more broadly across co-active facial areas (Rieger, 2024). This shift away from narrowly localized cues toward multi-region patterns is consistent with the objective of enforcing AU co-activation structure and illustrates how training constraints can alter not only performance but also how models process faces.

Experimental limitations: First, the evaluations in this chapter are computational; they intentionally exclude user studies. While this enables controlled, large-scale comparisons across models and explainers, assessing whether RelEx outputs are useful to clinicians or psychologists remains an important next step. Second, RelEx builds on post-hoc saliency maps (e.g., LRP, Grad-CAM). Known concerns about saliency faithfulness and robustness (Adebayo et al., 2018; Ali et al., 2023; Seth and Sankarapu, 2025) do not prevent RelExs use, since RelEx is explainer-agnostic. This motivates complementary diagnostics to validate the chosen explainer before auditing model plausibility.

Scope: RelEx provides a practical, extensible methodology for auditing whether models base decisions on physiologically meaningful facial regions—information that standard metrics overlook. By converting qualitative heatmaps into quantitative, AU-specific alignment scores, RelEx supports model selection, comparison across training regimes (e.g., with vs. without CorrLoss), and systematic monitoring for shortcut learning. The approach is architecture-agnostic and transferable to other domains with localized expert priors (e.g., medical imaging or industrial inspec-

tion), contributing to more trustworthy and auditable AI in settings where evidence location matters as much as prediction performance.

Conclusion and Outlook

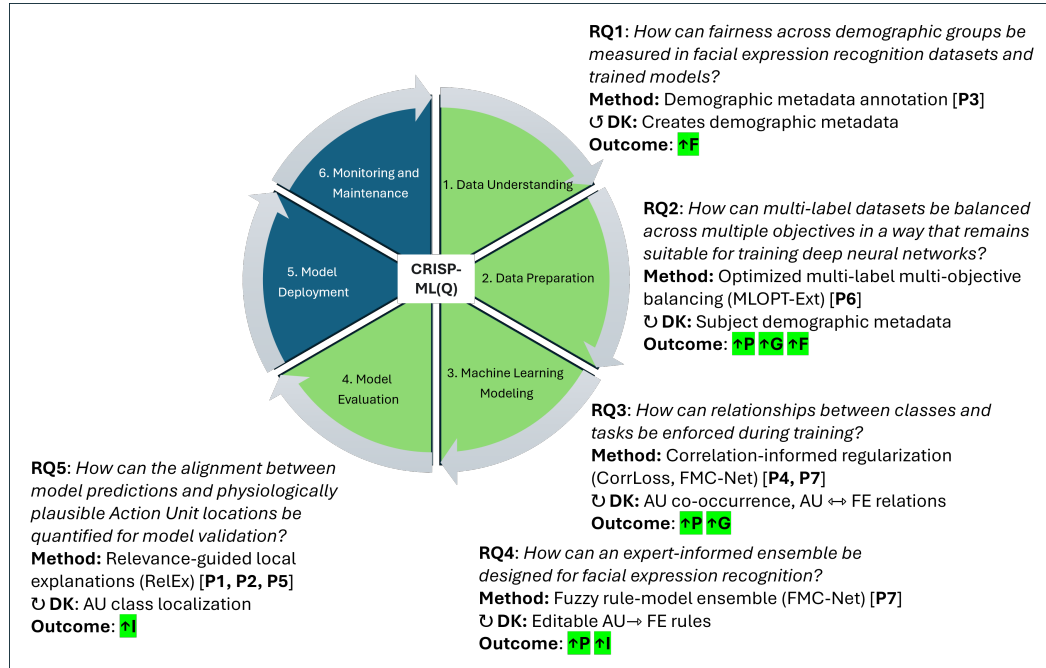


Fig. 9.1.: Overview of thesis contributions mapped to CRISP-ML(Q). Each research question (RQ) is shown with its method, associated publication(s) (P), and the domain knowledge (DK) (created ∪, injected ∩). Arrows indicate improvements per RQ: Performance (P), Interpretability (I), Robustness (R), Generalizability (G), Fairness (F).

9.1 Summary of Key Findings

This thesis addresses the overarching research question: *How can domain knowledge be integrated into deep neural networks to improve robustness and interpretability in facial expression recognition?* The five research sub-questions (RQ1–RQ5) are each treated in their respective chapters, which conclude with chapter-level summaries and discussions of findings and limitations.

By injecting domain knowledge—demographic metadata, AU co-occurrence structure, AU–FE task relations, expert logic rules, and AU localization—at multiple stages of the ML lifecycle (Fig. 9.1), this work introduces and empirically validates methods that collectively answer the overarching question. Together, these methods

improve generalization and demographic fairness (robustness), transparency (interpretability), and overall predictive performance, in line with the term boundaries introduced in Chapter 1. The main contributions and validated outcomes are:

- **RQ1 – Dataset/model fairness evaluation with subject metadata annotation.** Systematic annotation of subject metadata (ethnicity, age, gender, occlusions) exposed substantial demographic imbalance in widely used FER/AU datasets and corresponding performance disparities in state-of-the-art models ([P3] (Pahl et al., 2022)). This establishes a quantitative baseline for fairness assessment and motivates subsequent mitigation.
- **RQ2 – Multi-label and demographic balancing with MLOPT(-Ext).** MLOPT optimizes unique label-set frequencies under repetition bounds; MLOPT-Ext jointly balances task labels and demographics via a scalarized objective ([P6] (Rieger et al., 2020b)). Across benchmarks, these methods reduce imbalance, remain operational under realistic artefacts, and improve downstream performance and demographic parity.
- **RQ3 – Correlation-informed training (CorrLoss) and task constraints.** CorrLoss injects AU–AU structure as a differentiable regularizer and improves cross-dataset AU detection ([P4] (Rieger et al., 2022)). In FMC-Net, the idea is generalized to enforce AU↔FE consistency during multi-task training, reducing negative transfer and yielding stronger FER while preserving AU performance ([P7] (Rieger et al., 2025)).
- **RQ4 – Expert-informed, interpretable inference via a rule–model ensemble.** An editable fuzzy AU→FE pathway is ensembled with the neural FE model head to combine complementary errors and provide a disagreement-based risk signal for selective prediction and monitoring. This improves transparency and practical robustness without retraining when rules are adapted ([P7] (Rieger et al., 2025)).
- **RQ5 – Relevance-Guided Local Explanations (RelEx) for model audit.** RelEx is a post-hoc, model- and explainer-agnostic, domain-grounded audit that converts saliency maps into AU-specific plausibility scores using expert facial regions ([P5] (Rieger et al., 2020a), [P2] (Finzel et al., 2023)). It reveals right-for-the-wrong-reasons behavior in strong baselines and characterizes how CorrLoss reshapes model features toward distributed, co-active facial patterns ([P1] (Rieger, 2024)).

9.2 Discussion and Outlook

This section provides a high-level discussion and outlook; more detailed discussions appear in the respective method chapters.

Context and trade-offs: Robustness (Section 1.1.1) and interpretability (Section 1.1.2) are not single metrics but context-dependent goals that interact with each other and with accuracy through well-known trade-offs. Regularization can reduce variance and improve out-of-distribution behavior, yet it also shifts bias and may depress performance on particular subgroups or domains if the prior is misaligned. In this thesis, interpretability is operationalized along two axes at inference time: (1) structural transparency via rule-based ensembles, and (2) domain-plausible explanation maps via RelEx. Aligning inference with human-defined structure (e.g., fuzzy rules or AU localization) improves transparency and controllability, but the quality of this signal is bounded by the quality, completeness, and cultural specificity of the underlying expert knowledge. A practical path forward is therefore not a single best setting but an evaluation-driven design loop that balances accuracy, robustness, fairness, and transparency according to application risk.

Methods across the pipeline: The proposed methods intervene at multiple stages of the ML lifecycle. Dataset balancing (MLOPT-Ext) operationalizes fairness and coverage targets before training; correlation-informed objectives (CorrLoss/FMC-Net) inject structural priors during training; and the rule-model ensemble with disagreement provides a deployable reliability cue at inference. Finally, RelEx turns post-hoc explanations into task-aware, quantifiable audits, enabling comparisons that go beyond performance metrics. Together, these components demonstrate that domain knowledge (Section 1.2) can be encoded as constraints, priors, and audits that make performance gains more stable and more accountable.

Overall limitations: While the methods in this thesis show consistent improvements in robustness and interpretability, they also highlight several structural limitations that delineate the scope of the approach. First, there is a dependence on the quality of the domain knowledge, so components that encode priors (balancing targets, correlation structures, rule sets, AU regions) benefit from expert input, but this input is constrained by its completeness and potential bias. Second, bias-variance trade-offs remain: correlation-based regularization and dataset rebalancing can improve stability and generalization on average, yet may reduce performance in subsets where priors or sampling weights are misaligned. Third, RelEx is explainer-agnostic but relies on post-hoc saliency whose faithfulness can vary. Its scores should therefore be interpreted as plausibility indicators rather than causal proof. These limitations do not invalidate the approach, but they motivate complementary diagnostics (e.g., for explainer faithfulness) and context-aware deployment.

Outlook: Building on these findings, future work can extend this line of research in several directions. One avenue is to design richer priors, for example by incorporating temporal and contextual structure, modeling finer-grained AU compositions, and developing broader rule sets for nuanced or compound affective states. A second avenue is user-centered validation: pairing computational audits such as RelEx with expert studies to assess usefulness, trust, and editability, and co-designing interfaces for inspecting and modifying rules, targets, and constraints. Such studies would complement the computational perspective of this thesis with direct evidence of human interpretability and practical utility.

Broader impact: Although the methods were developed for FER, the proposed pattern (data balancing for coverage and fairness, correlation-aware training, rule–model ensembles for interpretability and deployable risk signals, domain-grounded audits) transfers to other multi-label or hierarchical domains (e.g., medical imaging, behavioral monitoring, industrial inspection). This aligns with principles of trustworthy AI and regulatory expectations (e.g., EU AI Act) by making model behavior measurably robust and transparent, while retaining the flexibility to adapt to differing risk profiles and stakeholder needs.

In sum, the thesis demonstrates that integrating domain knowledge across the pipeline yields systems that are not only more accurate on average, but also more stable under shift, more inspectable at the point of decision, and more amenable to governance.

Bibliography

- Adebayo, Julius, Justin Gilmer, Michael Muelly, et al. (2018). “Sanity checks for saliency maps”. In: *Advances in neural information processing systems* 31 (cit. on pp. 26, 74).
- Ali, Sajid, Tamer Abuhmed, Shaker El-Sappagh, et al. (2023). “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence”. In: *Information fusion* 99, p. 101805 (cit. on pp. 70, 74).
- Alipour, Mina, Mahyar T Moghaddam, Karthik Vaidhyanathan, and Mikkel Baun Kjærgaard (2023). “Emoticontrol: Emotions-based Control of User-Interfaces Adaptations”. In: *Proceedings of the ACM on Human-Computer Interaction* 7.EICS, pp. 1–29 (cit. on p. 56).
- Alvarez Melis, David and Tommi Jaakkola (2018). “Towards robust interpretability with self-explaining neural networks”. In: *Advances in neural information processing systems* 31 (cit. on p. 70).
- Atmaja, Bagus Tris and Masato Akagi (2021). “Evaluation of Error-and Correlation-Based Loss Functions for Multitask Learning Dimensional Speech Emotion Recognition”. In: *Journal of Physics: Conference Series*. Vol. 1896. 1. IOP Publishing, p. 012004 (cit. on p. 56).
- Bach, Sebastian, Alexander Binder, Grégoire Montavon, et al. (2015). “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation”. In: *PloS one* 10.7, e0130140 (cit. on pp. 26, 70, 71).
- Badreddine, Samy, Artur d’Avila Garcez, Luciano Serafini, and Michael Spranger (2022). “Logic Tensor Networks”. In: *Artificial Intelligence* 303, p. 103649 (cit. on pp. 64, 65).
- Baltrusaitis, Tadas, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency (2018). “OpenFace 2.0: Facial Behavior Analysis Toolkit”. In: *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pp. 59–66. DOI: 10.1109/FG.2018.00019 (cit. on p. 40).
- Bayer, Sarah, Henner Gimpel, and Moritz Markgraf (2022). “The role of domain expertise in trusting and following explainable AI decision support systems”. In: *Journal of Decision Systems* 32.1, pp. 110–138 (cit. on p. 5).
- Boutell, Matthew R, Jiebo Luo, Xipeng Shen, and Christopher M Brown (2009). “Learning multi-label scene classification”. In: *Pattern Recognition* 42.8, pp. 1756–1768 (cit. on p. 95).
- Bradski, G (2000). “The OpenCV library”. In: *Dr. Dobb’s Journal of Software Tools* (cit. on pp. 34, 35).
- Bruckert, Sebastian, Bettina Finzel, and Ute Schmid (2020). “The next generation of medical decision support: A roadmap toward transparent expert companions”. In: *Frontiers in artificial intelligence* 3, p. 507973 (cit. on p. 70).

- Charte, Francisco, Antonio Rivera, María José del Jesus, and Francisco Herrera (2013). “A first approach to deal with imbalance in multi-label datasets”. In: *International conference on hybrid artificial intelligence systems*. Springer, pp. 150–160 (cit. on p. 45).
- Charte, Francisco, Antonio J Rivera, Francisco Herrera, and María J del Jesus (2016). “The R Ultimate Multilabel Dataset Repository”. In: *International Conference on Hybrid Artificial Intelligence Systems*. Springer, Cham, pp. 487–499 (cit. on p. 95).
- Charte, Francisco, Antonio J Rivera, María J del Jesus, and Francisco Herrera (2015a). “Addressing imbalance in multilabel classification: Measures and random resampling algorithms”. In: *Neurocomputing* 163, pp. 3–16 (cit. on pp. 25, 45).
- (2015b). “MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation”. In: *Knowledge-Based Systems* 89, pp. 385–397 (cit. on p. 45).
- (2019). “REMEDIAL-HwR: Tackling multilabel imbalance through label decoupling and data resampling hybridization”. In: *Neurocomputing* 326, pp. 110–122 (cit. on p. 45).
- Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer (2002). “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16, pp. 321–357 (cit. on p. 44).
- Chen, Zhenpeng, Jie M Zhang, Max Hort, Mark Harman, and Federica Sarro (2024). “Fairness testing: A comprehensive survey and analysis of trends”. In: *ACM Transactions on Software Engineering and Methodology* 33.5, pp. 1–59 (cit. on p. 105).
- Conn, Andrew R, Nicholas IM Gould, and Philippe L Toint (2000). *Trust region methods*. SIAM (cit. on p. 50).
- Corneanu, Ciprian, Meysam Madadi, and Sergio Escalera (2018). “Deep structure inference network for facial action unit recognition”. In: *Proceedings of the european conference on computer vision (ECCV)*, pp. 298–313 (cit. on p. 56).
- Craig, Kenneth D, Kenneth M Prkachin, and Ruth VE Grunau (1992). “The facial expression of pain”. In: *Handbook of pain assessment* 2, pp. 257–276 (cit. on pp. 16, 18).
- Cui, Zijun, Tengfei Song, Yuru Wang, and Qiang Ji (2020). “Knowledge Augmented Deep Neural Networks for Joint Facial Expression and Action Unit Recognition”. In: *Advances in Neural Information Processing Systems* 33, pp. 14338–14349 (cit. on p. 56).
- Darwin, Charles Robert (1872). *The Expression of the Emotions in Man and Animals*. 1st. London: John Murray (cit. on p. 16).
- Dash, Tirtharaj, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan (2022). “A Review of Some Techniques for Inclusion of Domain-Knowledge Into Deep Neural Networks”. In: *Scientific Reports* 12.1, p. 1040 (cit. on pp. 5, 56).
- Dastin, Jeffrey (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. [Online; accessed 05-January-2022]. URL: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G> (cit. on p. 37).
- Dehshibi, Mohammad Mahdi and David Masip (2024). “BEE-NET: A deep neural network to identify in-the-wild Bodily Expression of Emotions”. In: *arXiv preprint arXiv:2402.13955* (cit. on p. 3).

- Deng, Didan, Zhaokang Chen, and Bertram E Shi (2020). “Multitask emotion recognition with incomplete labels”. In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, pp. 592–599 (cit. on p. 40).
- Deng, Jia, Wei Dong, Richard Socher, et al. (2009). “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255 (cit. on p. 22).
- Dhall, Abhinav, Roland Goecke, Simon Lucey, Tom Gedeon, et al. (2012). “Collecting large, richly annotated facial-expression databases from movies”. In: *IEEE multimedia* 19.3, p. 34 (cit. on p. 20).
- Ekman, Paul (1992). “An argument for basic emotions”. In: *Cognition & emotion* 6.3-4, pp. 169–200 (cit. on p. 16).
- Ekman, Paul and Daniel Cordaro (2011). “What is meant by calling emotions basic”. In: *Emotion review* 3.4, pp. 364–370 (cit. on p. 16).
- Ekman, Paul and Wallace V Friesen (1971). “Constants across cultures in the face and emotion.” In: *Journal of personality and social psychology* 17.2, p. 124 (cit. on p. 16).
- (1978). *Facial Action Coding Systems*. Consulting Psychologists Press (cit. on pp. xix, 3, 14, 15, 18, 56).
- Ekman, Paul, Wallace V Friesen, and Joseph C Hager (2002). “Facial action coding system: The manual on CD ROM”. In: *A Human Face, Salt Lake City*, pp. 77–254 (cit. on pp. xix, 14, 15, 18, 20).
- Fabian Benitez-Quiroz, C, Ramprakash Srinivasan, and Aleix M Martinez (2016). “Emotionet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild”. In: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5562–5570 (cit. on p. 33).
- Finzel, Bettina, **Ines Rieger**, Simon Kuhn, and Ute Schmid (2023). “Domain-Specific Evaluation of Visual Explanations for Application-Grounded Facial Expression Recognition”. In: *Machine Learning and Knowledge Extraction. CD-MAKE 2023. Lecture Notes in Computer Science*. Springer, pp. 31–44. DOI: [10.1007/978-3-031-40837-3_3](https://doi.org/10.1007/978-3-031-40837-3_3) (cit. on pp. xi, 8, 9, 69, 70, 72, 74, 78, 104).
- Friesen, Wallace V, Paul Ekman, et al. (1983). “Emfacs-7: Emotional Facial Action Coding System”. In: (cit. on pp. xix, 18, 19, 56).
- Gamage, Gihan, Daswin De Silva, Nishan Mills, Daminda Alahakoon, and Milos Manic (2024). “Emotion AWARE: an artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis”. In: *Journal of Big Data* 11.1, p. 93 (cit. on p. 56).
- Gavrilescu, Mihai and Nicolae Vizireanu (2019). “Predicting Depression, Anxiety, and Stress Levels From Videos Using the Facial Action Coding System”. In: *Sensors* 19.17, p. 3693 (cit. on p. 68).
- Gawlikowski, Jakob, Cedrique Rovile Njietcheu Tassi, Mohsin Ali, et al. (2023). “A survey of uncertainty in deep neural networks”. In: *Artificial Intelligence Review* 56.Suppl 1, pp. 1513–1589 (cit. on p. 64).
- Giorgio, Andrea de, Gabriele Cola, and Lihui Wang (2023). “Systematic review of class imbalance problems in manufacturing”. In: *Journal of Manufacturing Systems* 71, pp. 620–644 (cit. on pp. 43, 44).

- Girard, Jeffrey M., Wen-Sheng Chu, László A. Jeni, and Jeffrey F. Cohn (2017). “Sayette Group Formation Task (GFT) Spontaneous Facial Expression Database”. In: *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pp. 581–588. DOI: [10.1109/FG.2017.144](https://doi.org/10.1109/FG.2017.144) (cit. on pp. [6](#), [33](#)).
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press (cit. on p. [22](#)).
- Guidotti, Riccardo (2024). “Counterfactual explanations and how to find them: literature review and benchmarking”. In: *Data Mining and Knowledge Discovery* 38.5, pp. 2770–2824 (cit. on p. [70](#)).
- Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, et al. (2018). “A survey of methods for explaining black box models”. In: *ACM computing surveys (CSUR)* 51.5, pp. 1–42 (cit. on p. [4](#)).
- Gupta, Palak, Anmol Varshney, Mohammad Rafeek Khan, et al. (2023). “Unbalanced credit card fraud detection data: a machine learning-oriented comparative study of balancing techniques”. In: *Procedia Computer Science* 218, pp. 2575–2584 (cit. on p. [43](#)).
- Hájek, Petr, Lluís Godó, and Francesc Esteva (1996). “A Complete Many-Valued Logic With Product-Conjunction”. In: *Archive for Mathematical Logic* 35, pp. 191–208 (cit. on pp. [64](#), [65](#)).
- Han, ByungOk, Cheol-Hwan Yoo, Ho-Won Kim, Jang-Hee Yoo, and Jinhyeok Jang (2023). “Deep emotion change detection via facial expression analysis”. In: *Neurocomputing* 549, p. 126439 (cit. on p. [20](#)).
- Hassan, Teena, Dominik Seuß, Johannes Wollenberg, et al. (2019). “Automatic detection of pain from facial expressions: a survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.6, pp. 1815–1831 (cit. on p. [56](#)).
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (cit. on p. [22](#)).
- Holzinger, Andreas (2016). “Interactive machine learning for health informatics: when do we need the human-in-the-loop?” In: *Brain Informatics* 3.2, pp. 119–131 (cit. on p. [56](#)).
- Kamiran, Faisal and Toon Calders (2012). “Data preprocessing techniques for classification without discrimination”. In: *Knowledge and information systems* 33.1, pp. 1–33 (cit. on p. [46](#)).
- Karkkainen, Kimmo and Jungseock Joo (2021). “Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation”. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558 (cit. on pp. [6](#), [21](#), [37](#)).
- Kingma, Diederik P (2014). “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (cit. on p. [22](#)).
- Kollias, Dimitrios, Attila Schulc, Elnar Hajiye, and Stefanos Zafeiriou (2020). “Analysing affective behavior in the first abaw 2020 competition”. In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, pp. 637–643 (cit. on pp. [xii](#), [9](#), [29](#), [40](#), [53](#), [108](#)).

- Kollias, Dimitrios, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou (2023). “Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5888–5897 (cit. on p. 31).
- Kollias, Dimitrios, Panagiotis Tzirakis, Mihalios A Nicolaou, et al. (2019). “Deep Affect Prediction In-the-wild: Aff-wild Database and Challenge, Deep Architectures, and Beyond”. In: *International Journal of Computer Vision*, pp. 1–23 (cit. on p. 31).
- Kollias, Dimitrios and Stefanos Zafeiriou (2019). “Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace”. In: *arXiv preprint arXiv:1910.04855* (cit. on p. 31).
- Krippendorff, Klaus (2018). *Content analysis: An introduction to its methodology*. Sage publications (cit. on pp. 25, 39).
- (2004). “Reliability in content analysis: Some common misconceptions and recommendations”. In: *Human communication research* 30.3, pp. 411–433 (cit. on p. 25).
- Küblbeck, Christian and Andreas Ernst (2006). “Face detection and tracking in video sequences using the modifiedcensus transformation”. In: *Image and Vision Computing* 24.6, pp. 564–572 (cit. on pp. 34, 35).
- Kunz, Miriam and Stefan Lautenbacher (2014). “The Faces of Pain: A Cluster Analysis of Individual Differences in Facial Activity Patterns of Pain”. In: *European Journal of Pain* 18.6, pp. 813–823 (cit. on pp. xix, 16, 19, 56, 68).
- Kunz, Miriam, Doris Meixner, and Stefan Lautenbacher (2019). “Facial muscle movements encoding pain—a systematic review”. In: *Pain* 160.3, pp. 535–549 (cit. on p. 20).
- Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (2017). “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: *Advances in neural information processing systems* 30 (cit. on p. 64).
- Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, et al. (2019). “Unmasking Clever Hans predictors and assessing what machines really learn”. In: *Nature communications* 10.1, p. 1096 (cit. on p. 70).
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). “Deep learning”. In: *nature* 521.7553, pp. 436–444 (cit. on pp. 3, 21).
- Leong, Sze Chit, Yuk Ming Tang, Chung Hin Lai, and CKM Lee (2023). “Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing”. In: *Computer Science Review* 48, p. 100545 (cit. on pp. 3, 16, 21).
- Li, Guanbin, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin (2019). “Semantic Relationships Guided Representation Learning for Facial Action Unit Recognition”. In: *Proc. of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 8594–8601 (cit. on p. 56).
- Li, Shan and Weihong Deng (2020). “A deeper look at facial expression dataset bias”. In: *IEEE Transactions on Affective Computing* 13.2, pp. 881–893 (cit. on pp. 3, 4).
- Li, Ximan, Weihong Deng, Shan Li, and Yong Li (2023). “Compound Expression Recognition In-the-Wild With AU-Assisted Meta Multi-Task Learning”. In: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5734–5743 (cit. on pp. 3, 16, 67).

- Li, Yong and Shiguang Shan (2023). “Meta Auxiliary Learning for Facial Action Unit Detection”. In: *IEEE Transactions on Affective Computing* 14.03, pp. 2526–2538 (cit. on p. 56).
- Liang, Chen, Wenguan Wang, Jiaxu Miao, and Yi Yang (2023). “Logic-induced diagnostic reasoning for semi-supervised semantic segmentation”. In: *Proc. of IEEE/CVF Int. Conf on Computer Vision*, pp. 16197–16208 (cit. on p. 64).
- Liu, Tianci, Haoyu Wang, Yaqing Wang, et al. (2023). “SimFair: a unified framework for fairness-aware multi-label classification”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 12, pp. 14338–14346 (cit. on p. 46).
- Lucey, Patrick, Jeffrey F. Cohn, Takeo Kanade, et al. (2010). “The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-specified Expression”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101. DOI: [10.1109/CVPRW.2010.5543262](https://doi.org/10.1109/CVPRW.2010.5543262) (cit. on p. 32).
- Lucey, Patrick, Jeffrey F. Cohn, Kenneth M. Prkachin, Patricia E. Solomon, and Iain Matthews (2011). “Painful Data: The UNBC-McMaster Shoulder Pain Expression Archive Database”. In: *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pp. 57–64. DOI: [10.1109/FG.2011.5771462](https://doi.org/10.1109/FG.2011.5771462) (cit. on p. 34).
- Mavadati, Mohammad, Peyton Sanger, and Mohammad H Mahoor (2016). “Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8 (cit. on p. 32).
- Mavadati, S. Mohammad, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn (2013). “DISFA: A Spontaneous Facial Action Intensity Database”. In: *IEEE Transactions on Affective Computing* 4.2, pp. 151–160. DOI: [10.1109/T-AFFC.2013.4](https://doi.org/10.1109/T-AFFC.2013.4) (cit. on p. 32).
- Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan (2021). “A survey on bias and fairness in machine learning”. In: *ACM computing surveys (CSUR)* 54.6, pp. 1–35 (cit. on p. 21).
- Nourivandi, Tara, Saurabh Hinduja, Shivam Srivastava, Jeffrey F Cohn, and Shaun Canavan (2024). “Mitigating Class Imbalance for Facial Expression Recognition Using SMOTE on Deep Features”. In: *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, pp. 1–5 (cit. on pp. 3, 4, 6, 21, 37, 44).
- Oksuz, Kemal, Baris Can Cam, Sinan Kalkan, and Emre Akbas (2021). “Imbalance Problems in Object Detection: A Review”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 43.10, pp. 3388–3415 (cit. on p. 44).
- Pahl, Jaspar, Ines Rieger, and Dominik Seuss (2020a). “Multi-label class balancing algorithm for action unit detection”. In: *arXiv preprint arXiv:2002.03238* (cit. on p. 108).
- Pahl, Jaspar, **Ines Rieger**, Anna Möller, Thomas Wittenberg, and Ute Schmid (2022). “Female, white, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces”. In: *Proc. of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, 973–987. DOI: [10.1145/3531146.3533159](https://doi.org/10.1145/3531146.3533159) (cit. on pp. xi, xvii, xix, 4, 6, 8, 30, 37–40, 44, 78, 105).

- Pahl, Jaspar, **Ines Rieger**, and Dominik Seuss (2020b). “Multi-label Learning with Missing Values using Combined Facial Action Unit Datasets”. In: *The 37th International Conference on Machine Learning (ICML), Workshop on the Art of Learning with Missing Values (Artemiss)*. URL: https://openreview.net/forum?id=_wVkmK7wBhW (cit. on p. 104).
- Pal, Madhura, J D Greene, and Stephen M Weiss (2005). “A knowledge-based approach to prediction of protein functions”. In: *Fifth IEEE International Conference on Data Mining (ICDM’05)*. IEEE, 8–pp (cit. on p. 95).
- Panigutti, Cecilia, Ronan Hamon, Isabelle Hupont, et al. (2023). “The role of explainable AI in the context of the AI Act”. In: *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 1139–1150 (cit. on p. 3).
- Parham Salar (2020). *Corel-5k Dataset*. <https://www.kaggle.com/datasets/parhamsalar/corel5k>. Accessed: 2025-06-09 (cit. on p. 95).
- Picard, Rosalind W (2000). *Affective computing*. MIT press (cit. on p. 3).
- Pons, Gerard and David Masip (2018). “Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis”. In: *IEEE Trans. on Aff. Comp.* 9.03, pp. 343–350 (cit. on p. 64).
- Posner, Jonathan, James A Russell, and Bradley S Peterson (2005). “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology”. In: *Development and psychopathology* 17.3, pp. 715–734 (cit. on p. 16).
- Rabold, Johannes, Gesina Schwalbe, and Ute Schmid (2020). “Expressive explanations of DNNs by combining concept analysis with ILP”. In: *KI 2020: Advances in Artificial Intelligence: 43rd German Conference on AI, Bamberg, Germany, September 21–25, 2020, Proceedings* 43. Springer, pp. 148–162 (cit. on p. 70).
- Rabold, Johannes, Michael Siebers, and Ute Schmid (2022). “Generating contrastive explanations for inductive logic programming based on a near miss approach”. In: *Machine Learning* 111.5, pp. 1799–1820 (cit. on p. 70).
- Ramis, Silvia, Jose M Buades, Francisco J Perales, and Cristina Manresa-Yee (2022). “A novel approach to cross dataset studies in facial expression recognition”. In: *Multimedia Tools and Applications* 81.27, pp. 39507–39544 (cit. on p. 106).
- Rezvani, Salim and Xizhao Wang (2023). “A broad review on class imbalance learning techniques”. In: *Applied Soft Computing*, p. 110415 (cit. on p. 44).
- Rieger, Ines, Thomas Hauenstein, Sebastian Hettenkofer, and Jens-Uwe Garbas (2019). “Towards real-time head pose estimation: Exploring parameter-reduced residual networks on in-the-wild datasets”. In: *Advances and Trends in Artificial Intelligence. From Theory to Practice: 32nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2019*. Springer, pp. 123–134 (cit. on pp. 22, 107).
- Ringeval, Fabien, Björn Schuller, Michel Valstar, et al. (2015). “AV+EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data”. In: *Proc. of the 5th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–8 (cit. on p. 56).

- Robinson, Joseph P, Can Qin, Yann Henon, Samson Timoner, and Yun Fu (2023). “Balancing biases and preserving privacy on balanced faces in the wild”. In: *IEEE Transactions on Image Processing* 32, pp. 4365–4377 (cit. on p. 37).
- Ruiz, Adria, Joost Van de Weijer, and Xavier Binefa (2015). “From Emotions to Action Units with Hidden and Semi-hidden-task Learning”. In: *Proc. of the IEEE International Conference on Computer Vision*, pp. 3703–3711 (cit. on p. 63).
- Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller (2017). “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models”. In: *arXiv preprint arXiv:1708.08296* (cit. on p. 70).
- Sarmah, Devashish and Nityananda Das (2021). “Emotions: A dataset for emotion analysis in text”. In: *arXiv preprint arXiv:2105.02324* (cit. on p. 95).
- Schmid, Ute (2023). “Trustworthy Artificial Intelligence: Comprehensible, Transparent and Correctable”. In: *Introduction to Digital Humanism: A Textbook*. Springer (cit. on p. 8).
- (2024). “Trustworthy Artificial Intelligence: Comprehensible, Transparent and Correctable”. In: *Hannes Werthner · Carlo Ghezzi · Jeff Kramer · Julian Nida-Rümelin · Bashar Nuseibeh · Erich Prem*, p. 151 (cit. on pp. 4, 70).
- Schwalbe, Gesina, Christian Wirth, and Ute Schmid (2022). “Enabling Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings”. In: *arXiv preprint arXiv:2201.00572* (cit. on p. 64).
- Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, et al. (2017). “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proc. of ICCV* (cit. on pp. 26, 70, 71).
- Seth, Pratinav and Vinay Kumar Sankarapu (2025). “Bridging the Gap in XAI-Why Reliable Metrics Matter for Explainability and Compliance”. In: *arXiv preprint arXiv:2502.04695* (cit. on pp. 70, 74).
- Seuss, Dominik, Anja Dieckmann, Teena Hassan, et al. (2019). “Emotion Expression from Different Angles: A Video Database for Facial Expressions of Actors Shot by a Camera Array”. In: *Proc. of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 35–41. DOI: [10.1109/ACII.2019.8925458](https://doi.org/10.1109/ACII.2019.8925458) (cit. on pp. 6, 29, 31, 62).
- Seuss, Dominik, Teena Hassan, Anja Dieckmann, et al. (2023). “Automatic Estimation of Action Unit Intensities and Inference of Emotional Appraisals”. In: *IEEE Transactions on Affective Computing* 14.2, pp. 1188–1200 (cit. on pp. 63, 67).
- Shao, Zhiwen, Yong Zhou, Feiran Li, Hancheng Zhu, and Bing Liu (2024). “Joint facial action unit recognition and self-supervised optical flow estimation”. In: *Pattern Recognition Letters* 181, pp. 70–76 (cit. on p. 56).
- Shneiderman, Ben (2022). *Human-centered AI*. Oxford, UK: Oxford University Press (cit. on pp. 5, 56).
- Siebers, Michael, Ute Schmid, Dominik Seuß, Miriam Kunz, and Stefan Lautenbacher (2016). “Characterizing Facial Expressions by Grammars of Action Unit Sequences—a First Investigation Using ABL”. In: *Information Sciences* 329, pp. 866–875 (cit. on p. 68).
- Simonyan, Karen and Andrew Zisserman (2015). “Very deep convolutional networks for large-scale image recognition”. In: *ICLR* (cit. on p. 22).

- Sivakumar, Jayanth, Karthik Ramamurthy, Menaka Radhakrishnan, and Daehan Won (2022). “Synthetic sampling from small datasets: A modified mega-trend diffusion approach using k-nearest neighbors”. In: *Knowledge-based systems* 236, p. 107687 (cit. on p. 45).
- Snoek, Cees G M, Marcel Worring, Jan-Mark van Gemert, Jan M Geusebroek, and Arnold W M Smeulders (2006). “The challenge problem for automated detection of 101 semantic concepts in multimedia”. In: *Proceedings of the 14th ACM international conference on Multimedia*, pp. 421–430 (cit. on p. 95).
- Song, Tengfei, Zijun Cui, Wenming Zheng, and Qiang Ji (2021a). “Hybrid message passing with performance-driven structures for facial action unit detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6267–6276 (cit. on p. 56).
- Song, Zeyu, Dongliang Chang, Zhanyu Ma, Xiaoxu Li, and Zheng-Hua Tan (2021b). “CC-Loss: Channel Correlation Loss for Image Classification”. In: *25th International Conference on Pattern Recognition (ICPR)*, pp. 7601–7608. DOI: [10.1109/ICPR48806.2021.9412069](https://doi.org/10.1109/ICPR48806.2021.9412069) (cit. on p. 56).
- Studer, Stefan, Thanh Binh Bui, Christian Drescher, et al. (2021). “Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology”. In: *Machine learning and knowledge extraction* 3.2, pp. 392–413 (cit. on p. 5).
- Rieger, Ines** (2024). “Investigating the Regularization of Deep Neural Networks for Affect Recognition With Relevance-Guided Local Explanations”. In: *Artificial Intelligence: Methodology, Systems, and Applications: 19th International Conference, AIMSA 2024*. Lecture Notes in Computer Science. Springer, pp. 122–127. DOI: [10.1007/978-3-031-81542-3_10](https://doi.org/10.1007/978-3-031-81542-3_10) (cit. on pp. xi, xvii, 8, 9, 69–71, 74, 78, 103).
- Rieger, Ines**, René Kollmann, Bettina Finzel, Dominik Seuss, and Ute Schmid (2020a). “Verifying Deep Learning-based Decisions for Facial Expression Recognition”. In: *Proc. of the 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pp. 139–144. URL: <https://www.esann.org/sites/default/files/proceedings/2020/ES2020-49.pdf> (cit. on pp. xi, 8, 9, 69, 70, 72, 74, 78, 104, 107).
- Rieger, Ines**, Jaspar Pahl, Bettina Finzel, and Ute Schmid (2022). “CorrLoss: Integrating Co-Occurrence Domain Knowledge for Affect Recognition”. In: *Proc. of the 26th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 798–804. DOI: [10.1109/ICPR56361.2022.9956319](https://doi.org/10.1109/ICPR56361.2022.9956319) (cit. on pp. xi, xvii, 7, 9, 55, 56, 62, 78, 106, 109).
- Rieger, Ines**, Jaspar Pahl, and Ute Schmid (2025). “FMC-Net: A Human-Guided Deep Learning Framework for Adaptable and Transparent Facial Expression Recognition in Real-World Scenarios”. In: *Applied Intelligence* 55.18. DOI: [10.1007/s10489-025-07017-9](https://doi.org/10.1007/s10489-025-07017-9) (cit. on pp. xii, xvii, xix, 7, 9, 19, 55, 56, 62–64, 78, 109).
- Rieger, Ines**, Jaspar Pahl, and Dominik Seuss (2020b). “Unique Class Group Based Multi-Label Balancing Optimizer for Action Unit Detection”. In: *Proc. of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG), Workshop on Affect Recognition in-the-wild: Uni/Multi-Modal Analysis and VA-AU-Expression*

- Challenges*. IEEE, pp. 619–623. DOI: [10.1109/FG47880.2020.00101](https://doi.org/10.1109/FG47880.2020.00101) (cit. on pp. [xii](#), [6](#), [8](#), [43](#), [45](#), [46](#), [53](#), [78](#), [108](#)).
- Tian, Y-I, Takeo Kanade, and Jeffrey F Cohn (2001). “Recognizing action units for facial expression analysis”. In: *IEEE Transactions on pattern analysis and machine intelligence* 23.2, pp. 97–115 (cit. on p. [19](#)).
- Turnbull, Douglas, Luke Barrington, David Torres, and Gert Lanckriet (2008). “Semantic annotation and retrieval of music and sound effects”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 16.2, pp. 467–476 (cit. on p. [95](#)).
- Valstar, Michel François and Maja Pantic (2006). “Biologically vs. logic inspired encoding of facial actions and emotions in video”. In: *2006 IEEE Int. Conf. on Multimedia and Expo*. IEEE, pp. 325–328 (cit. on p. [63](#)).
- Van Lent, Michael, William Fisher, and Michael Mancuso (2004). “An explainable artificial intelligence system for small-unit tactical behavior”. In: *Proc. of AAAI* (cit. on p. [4](#)).
- Velusamy, Sudha, Hariprasad Kannan, Balasubramanian Anand, Anshul Sharma, and Bilva Navathe (2011). “A Method to Infer Emotions from Facial Action Units”. In: *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2028–2031 (cit. on p. [63](#)).
- Vincent, James (2018). *Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech*. [Online; accessed 05-January-2022]. URL: <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai> (cit. on p. [37](#)).
- Watson, David S (2022). “Conceptual challenges for interpretable machine learning”. In: *Synthese* 200.2, p. 65 (cit. on pp. [4](#), [8](#)).
- Yang, Huiyuan, Lijun Yin, Yi Zhou, and Jiuxiang Gu (2021). “Exploiting Semantic Embedding and Visual Feature for Facial Action Unit Detection”. In: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10482–10491 (cit. on pp. [3](#), [56](#)).
- Yu, Jun, Zhongpeng Cai, Renda Li, et al. (2023). “Exploring large-scale unlabeled faces to enhance facial expression recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5803–5810 (cit. on p. [3](#)).
- Zafeiriou, Stefanos, Dimitrios Kollias, Mihalis A Nicolaou, et al. (2017). “Aff-wild: valence and arousal ‘In-the-Wild’ challenge”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 34–41 (cit. on pp. [29](#), [31](#)).
- Zhang, Aston, Zachary C Lipton, Mu Li, and Alexander J Smola (2023a). *Dive into deep learning*. Cambridge University Press (cit. on pp. [3](#), [21](#)).
- Zhang, Wei, Bowen Ma, Feng Qiu, and Yu Ding (2023b). “Multi-modal Facial Affective Analysis based on Masked Autoencoder”. In: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5792–5801 (cit. on p. [56](#)).
- Zhang, Wei, Feng Qiu, Chen Liu, et al. (2024a). “An Effective Ensemble Learning Framework for Affective Behaviour Analysis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4761–4772 (cit. on p. [64](#)).
- Zhang, Wei, Feng Qiu, Suzhen Wang, et al. (2022). “Transformer-based multimodal information fusion for facial expression analysis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2428–2437 (cit. on p. [56](#)).

- Zhang, Xiaowei, Bin Tang, Xiaowei Sun, et al. (2024b). “GenBase: A Nucleotide Sequence Database”. In: *Genomics, Proteomics & Bioinformatics* 22.3, qzae047 (cit. on p. 95).
- Zhang, Xing, Lijun Yin, Jeffrey F Cohn, et al. (2014). “Bp4d-Spontaneous: A High-Resolution Spontaneous 3D Dynamic Facial Expression Database”. In: *Image and Vision Computing* 32.10, pp. 692–706 (cit. on pp. 6, 32, 95).
- Zhang, Zheng, Jeff M Girard, Yue Wu, et al. (2016). “Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3438–3446 (cit. on p. 32).
- Zhao, Kaili, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang (2015). “Joint Patch and Multi-Label Learning for Facial Action Unit Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2207–2216 (cit. on p. 56).
- Zhao, Kaili, Wen-Sheng Chu, and Honggang Zhang (2016). “Deep region and multi-label learning for facial action unit detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3391–3399 (cit. on p. 70).

Part II

Publications of Dissertation

Additional Results

This section includes additional experimental results for the Chapter 5 that are currently unpublished but are intended for future publication.

My scientific contributions:

- Developing the experimental design for additional MLOPT experiments in collaboration with Jaspar Pahl.
- Developing the research concept and experimental design for MLOPT-Ext in collaboration with Jaspar Pahl. This extension improves upon the MLOPT class balancing algorithm by including the balancing of subject demographic metadata.

A.1 Overview of Tabular Benchmark Datasets

Table A.1 provides a brief overview with key characteristics of the datasets utilized in the additional experiments. The benchmark datasets used in these experiments cover a diverse range of domains for multi-label classification. These include cal500 (Turnbull et al., 2008) for music and audio analysis, corel5k (Parham Salar, 2020) and corel16k (Charte et al., 2016) as benchmark image datasets, an emotions (Sarmah and Das, 2021) dataset for text-based emotion analysis, and genbase (Zhang et al., 2024b) as a nucleotide sequence database. Additionally, the mediamill (Snoek et al., 2006) dataset focuses on semantic concept detection in videos, yeast (Pal et al., 2005) on protein function prediction from interaction networks, and scene (Boutell et al., 2009) for multi-label scene classification, and bp4d (Zhang et al., 2014) as a database for spontaneous 3D dynamic facial expressions. The multi-label multi-class distribution of these datasets is used for evaluating the balancing algorithm MLOPT(-Ext).

Tab. A.1.: Overview of datasets. MeanIR is the average of the imbalance ratios across all labels.

Dataset	Instances	Classes	Label-Sets	MeanIR
cal500	502	174	502	20.5778
corel5k	5000	374	3175	189.5676
corel16k	13,766	153	4803	34.1552
emotions	593	6	27	1.4781
genbase	662	27	32	37.3146
mediamill	43,907	101	3175	256.4047
yeast	2417	14	198	7.1968
scene	2407	6	15	1.2538
bp4d	1,46,346	15	1301	2.715

A.2 Comparison of MLOPT with Different Objective Functions

This section presents a comparative analysis of the MLOPT algorithm when optimized with different distance terms in the objective function: L1, L2, and Kullback-Leibler Divergence (KLDIV), under the parameters of $\lambda = 0$ and $\text{bounds} = 10$. Both L1 and L2 lead to consistent improvements over the original imbalance of the datasets (Table A.1), as reflected in reduced MeanIR values. In contrast, optimization with KLDIV results in higher MeanIR values — exceeding the original imbalance. This suggests that KLDIV is less effective, or even counterproductive, for the specific goal of balancing multi-label distributions within the MLOPT framework.

As detailed in Table A.2, using L2 consistently demonstrates superior performance in terms of MeanIR across all evaluated datasets. A lower MeanIR value indicates a more balanced label distribution (see Section 2.4.1). In most instances (cal500, corel5k, corel16k, genbase, mediamill, yeast), L2 achieves the lowest MeanIR. For the emotions and scene datasets, L2 ties with L1 for the lowest MeanIR value.

While the table also reports the final values of each objective function, these are only meaningful within the context of the respective optimization target. A lower value indicates better minimization of that specific objective function but does not allow for direct comparison across different ones. The key insight is the empirical effect of these optimizations: despite not explicitly minimizing MeanIR, using the L2 norm consistently leads to the most favorable MeanIR values. This indicates that L2 is particularly well-suited for promoting balanced label distributions in the multi-label setting and serves as a strong default choice for the optimization of MLOPT.

Tab. A.2.: Comparison of L1, L2, and Kullback-Leibler Divergence (KLDIV). Best values per dataset are in bold.

Dataset	L1		L2		KLDIV	
	Objective	MeanIR	Objective	MeanIR	Objective	MeanIR
cal500	9.8×10^{-2}	19.82	1.7×10^{-2}	17.32	1.0×10^{-1}	43.63
corel5k	7.0×10^{-3}	92.87	1.4×10^{-4}	78.64	1.9×10^{-2}	874.63
corel16k	8.0×10^{-3}	15.52	2.0×10^{-4}	8.28	3.0×10^{-2}	160.18
emotions	2.9×10^{-15}	1.00	1.7×10^{-15}	1.00	6.4×10^{-2}	1.84
genbase	3.4×10^{-2}	17.65	2.0×10^{-3}	4.61	8.1×10^{-2}	47.01
mediamill	3.3×10^{-2}	229.81	3.6×10^{-3}	171.80	3.2×10^{-2}	701.54
yeast	6.8×10^{-2}	4.87	8.0×10^{-3}	3.36	7.8×10^{-2}	25.25
scene	4.6×10^{-11}	1.00	3.3×10^{-15}	1.00	1.1×10^{-1}	6.34

A.3 Comparison of MLOPT with Competitive Approaches

This section details additional experimental results, benchmarking the MLOPT algorithm against two prominent multi-label oversampling balancing algorithms: LPROS and MLROS, as presented in Table 5.1. The findings are summarized in Section 5.3.1.

The comparison is structured around two main experimental setups, with MeanIR serving as the primary metric for evaluating the achieved label distribution balance. For fairness, MLOPT is optimized separately using the MeanIR and L2 objectives, as the competitive methods LPROS and MLROS are optimized solely for MeanIR. All MLOPT experiments are conducted with $\lambda = 0$. The first set of experiments, presented in Tables A.3 and A.4, explores the algorithms performance when growth is constrained by bounds. MLOPT utilizes these bounds to define the maximum permissible repetition of unique label-sets relative to their original count (e.g., up to 5 or 10 times their original count). In contrast, LPROS and MLROS employ a percentage increase as their constraint mechanism. LPROS oversamples instances based on whether they are below the mean. Consequently, this algorithm may encounter errors with already balanced datasets, such as cal500, where they are unable to operate. Similarly, in the mediamill dataset, some instances lack true labels, preventing the MLROS algorithm from functioning correctly.

The second set of experiments investigates performance when a percentage increase is set, as shown in Tables A.5 and A.6. Here, the percentage increase refers to the allowed growth of the original dataset for LPROS and MLROS. For MLOPT, the maximum repetition (bound) is calculated based on the maximum repetition observed for MLROS under the given percentage increase. The best MeanIR result

for each dataset across all algorithms in these tables is highlighted in bold. Across all experimental settings—both with bounded repetition and fixed percentage increases, MLOPT consistently achieves the best label distribution compared to LPROS and MLROS, as measured by MeanIR. When comparing the two objective functions used within MLOPT, the results show that the L2 objective tends to outperform direct MeanIR optimization in terms of achieving lower MeanIR values, particularly under less restrictive constraints, such as higher bounds (e.g., 10) or percentage increases (e.g., 20%). In these settings, where the algorithm has greater flexibility to adjust label-set frequencies, the L2 objective more effectively promotes a flatter and more uniform label distribution, which translates into better MeanIR scores.

Tab. A.3.: Comparison of competitive algorithms for all datasets with bounds = 5. Metric is MeanIR.

Datasets		MLOPT		LPROS	MLROS
Name	MeanIR	MeanIR	L2	MeanIR	MeanIR
cal500	20.5778	11.890	17.218	-	13.915
corel5k	189.5676	60.435	80.044	341.962	89.863
corel16k	34.1552	11.698	12.416	68.218	22.869
emotions	1.4781	1.022	1.000	1.214	1.392
genbase	37.3146	9.234	8.044	9.335	15.986
mediamill	256.4047	81.005	185.928	130.598	-
yeast	7.1968	2.705	3.919	3.184	3.605
scene	1.2538	1.019	1.000	1.635	1.208

Tab. A.4.: Comparison of competing algorithms for all datasets with bounds = 10. Metric is MeanIR.

Datasets		MLOPT		LPROS	MLROS
Name	MeanIR	MeanIR	L2	MeanIR	MeanIR
cal500	20.5778	9.179	17.318	-	12.473
corel5k	189.5676	32.312	78.640	341.962	89.719
corel16k	34.1552	7.927	8.281	68.218	22.934
emotions	1.4781	1.024	1.000	1.214	1.401
genbase	37.3146	5.447	4.614	9.335	15.986
mediamill	256.4047	52.324	171.798	130.598	-
yeast	7.1968	2.756	2.309	3.184	3.582
scene	1.2538	1.012	1.000	1.635	1.220

Tab. A.5.: Comparison of competing algorithms for all datasets with % increase = 10. Metric is MeanIR.

Datasets		MLOPT		LPROS	MLROS
Name	MeanIR	MeanIR	L2	MeanIR	MeanIR
cal500	20.5778	15.469	19.025	-	18.265
corel5k	189.5676	45.635	77.898	341.962	89.825
corel16k	34.1552	7.939	6.522	95.354	25.178
emotions	1.4781	1.014	1.000	1.696	1.438
genbase	37.3146	9.234	8.128	26.137	16.080
mediamill	256.4047	161.914	102.985	276.005	-
yeast	7.1968	2.029	1.322	4.148	3.751
scene	1.2538	1.012	1.000	2.178	1.208

Tab. A.6.: Comparison of competing algorithms for all datasets with % increase = 20. Metric is MeanIR.

Datasets		MLOPT		LPROS	MLROS
Name	MeanIR	MeanIR	L2	MeanIR	MeanIR
cal500	20.5778	13.267	18.377	-	16.787
corel5k	189.5676	41.203	78.640	341.962	90.270
corel16k	34.1552	11.205	5.964	95.354	22.982
emotions	1.4781	1.022	1.000	1.507	1.420
genbase	37.3146	9.169	8.044	17.350	16.079
mediamill	256.4047	163.411	103.476	194.770	-
yeast	7.1968	2.408	1.197	3.920	3.562
scene	1.2538	1.024	1.000	2.000	1.218

A.4 Evaluation of MLOPT-Ext on BP4D

This section reports experiments applying the extended balancing algorithm MLOPT-Ext on the BP4D dataset. The goal is to jointly mitigate demographic imbalance (metadata: ethnicity, age, gender) while maintaining a balanced distribution across selected AUs. An additional parameter, η , controls the relative weighting of metadata balancing versus AU balancing. The main findings are summarized in Section 5.3.3. For these experiments, a specific set of impactful AUs for facial expressions was utilized: 1, 2, 4, 6, 7, 12, 15, and 17. The dataset was optimized for combinations of η and λ (each in $[0, 1]$) under both L1 and L2 regularization. All optimization runs used bounds=10.

The primary outcomes are reported as standard deviations that quantify (1) variability across demographic groups (ethnicity, age, gender) and (2) variability across AUs. Lower values indicate more uniformity with respect to the corresponding attribute. In addition, MeanIR is reported as a summary indicator of overall class balance.

Table A.7 summarizes all configurations. Values exceeding the unoptimized baselines (given in the caption) are underlined to highlight settings that do not improve (or even degrade) the respective criterion. For readability, we additionally bold, within each row, the smaller value between the L1 and L2 results for each metric; this formatting is intended purely as a visual aid for per-metric comparison and should not be interpreted as a single jointly attainable configuration.

Overall, settings with $\eta = 0.0$ tend to yield little improvement for demographic dispersion. This is expected because $\eta = 0$ removes the explicit incentive to reduce disparities in ethnicity, age, and gender, and thus the optimization primarily targets AU-related balancing. Consequently, several demographic standard deviations remain near the baseline or become worse (underlined), even when small improvements in other criteria are observed.

Across the explored grid, increasing η generally encourages stronger reductions in metadata dispersion, but these gains may come with trade-offs in other objectives, most notably the AU standard deviation. This motivates treating the results as a multi-objective trade-off rather than expecting a single configuration to simultaneously minimize all reported quantities.

Empirically, L1 and L2 regularization exhibit different trade-off profiles. L1 frequently achieves very small dispersion values for age and gender (often reaching the numerical resolution shown in the table), and can also achieve strong reductions in ethnicity dispersion for larger η and λ . In contrast, L2 can yield competitive (and in some cases substantially lower) ethnicity dispersion at specific settings (e.g., $\eta = 1.0, \lambda = 0.0$), and achieves the lowest MeanIR observed in this grid at $\eta = 0.5, \lambda = 0.0$. These observations indicate that neither regularizer uniformly dominates across all objectives; instead, the preferable choice depends on which criteria are prioritized.

Given the competing nature of AU dispersion and metadata dispersion, a Pareto-front analysis could be used to summarize the trade-off: one can treat each (η, λ) setting (and each choice of regularizer) as a candidate solution and compare it under two objectives, e.g., minimizing σ_{AU} and a scalar metadata dispersion σ_{meta} (such as the mean or maximum over ethnicity, age, and gender). Under such a non-dominance criterion, multiple configurations would typically remain on the Pareto front, reflecting that improvements in one objective can require compromises in the other. Importantly, Table A.7 already indicates that strong improvements are not confined to a single setting: several configurations with $\eta > 0$ achieve substantially reduced metadata dispersion while also maintaining low AU dispersion and MeanIR values. This suggests that η primarily controls where the method operates along the AU–metadata trade-off curve, and that there exist multiple practically attractive operating points depending on the intended application priorities.

Tab. A.7.: MLOPT-Ext applied on BP4D dataset. Original (unoptimized) standard deviations are: AU: 0.28155, Ethnicity: 0.3322, Age: 0.1518, Gender: 0.1188. Unoptimized MeanIR: 2.715. Values worse than these original values are underlined; for each row, the better value between L1 and L2 for each attribute is **bolded**.

η	λ	L1					L2				
		AUs	Ethnicity	Age	Gender	MeanIR	AUs	Ethnicity	Age	Gender	MeanIR
0.0	0.0	0.1172	0.3363	0.1678	0.0860	<u>2.821</u>	0.0603	0.3312	<u>0.1574</u>	<u>0.1221</u>	2.735
0.0	0.1	0.0887	<u>0.3333</u>	<u>0.1654</u>	0.0950	<u>2.790</u>	0.2742	<u>0.3323</u>	0.1513	0.1167	2.698
0.0	0.2	0.0596	0.3315	<u>0.1586</u>	0.1012	2.734	0.2782	0.3322	0.1516	0.1178	2.706
0.0	0.3	0.0534	0.3309	<u>0.1553</u>	0.0991	2.712	0.2796	0.3322	0.1517	0.1183	2.709
0.0	0.4	0.1171	0.3361	<u>0.1708</u>	0.0888	2.834	0.2803	0.3322	0.1518	0.1185	2.711
0.0	0.5	0.1021	<u>0.3346</u>	<u>0.1686</u>	0.0930	<u>2.806</u>	0.2807	0.3322	0.1518	0.1186	2.712
0.0	1.0	0.2818	<u>0.3323</u>	<u>0.1519</u>	<u>0.1190</u>	2.715	0.2818	<u>0.3323</u>	<u>0.1519</u>	<u>0.1190</u>	2.715
0.1	0.0	0.0403	0.3280	0.0382	0.0001	2.364	0.0778	0.3023	0.0767	0.0565	2.170
0.1	0.1	0.0642	0.3283	0.0328	0.0000	2.365	0.2754	0.3320	0.1502	0.1161	2.691
0.1	0.2	0.0651	0.3283	0.0147	0.0000	2.366	0.2787	0.3321	0.1510	0.1175	2.703
0.1	0.3	0.1141	0.3263	0.0539	0.0000	2.328	0.2799	0.3321	0.1513	0.1180	2.707
0.1	0.4	0.1104	0.3259	0.0298	0.0000	2.309	0.2805	0.3321	0.1515	0.1183	2.710
0.1	0.5	0.1042	0.3231	0.0001	0.0000	2.258	0.2809	0.3322	0.1516	0.1184	2.711
0.2	0.0	0.0719	0.3212	0.0003	0.0000	2.222	0.0919	0.2798	0.0525	0.0334	2.014
0.2	0.1	0.0763	0.3184	0.0003	0.0000	2.189	0.2759	0.3317	0.1489	0.1151	2.684
0.2	0.2	0.0860	0.3164	0.0002	0.0000	2.169	0.2790	0.3320	0.1503	0.1170	2.699
0.2	0.3	0.0854	0.3110	0.0003	0.0000	2.138	0.2802	0.3320	0.1509	0.1177	2.705
0.2	0.4	0.0714	0.3001	0.0017	0.0017	2.083	0.2807	0.3321	0.1511	0.1180	2.708
0.2	0.5	0.1311	0.2954	0.0003	0.0000	2.036	0.2811	0.3321	0.1513	0.1182	2.710
0.3	0.0	0.0827	0.3082	0.0003	0.0000	2.121	0.1024	0.2623	0.0423	0.0215	1.942
0.3	0.1	0.1222	0.3063	0.0003	0.0001	2.084	0.2767	0.3315	0.1478	0.1144	2.678
0.3	0.2	0.1635	0.3122	0.0003	0.0001	2.131	0.2795	0.3318	0.1498	0.1167	2.697
0.3	0.3	0.1057	0.2804	0.0003	0.0002	1.967	0.2805	0.3320	0.1505	0.1175	2.703
0.3	0.4	0.1581	0.2861	0.0002	0.0000	2.012	0.2810	0.3320	0.1509	0.1179	2.707
0.3	0.5	0.1866	0.2776	0.0002	0.0001	1.994	0.2813	0.3321	0.1511	0.1182	2.709
0.4	0.0	0.1204	0.2961	0.0003	0.0000	2.041	0.1153	0.2500	0.0345	0.0156	1.896
0.4	0.1	0.1792	0.3076	0.0002	0.0000	2.119	0.2776	0.3313	0.1469	0.1138	2.673
0.4	0.2	0.1197	0.2633	0.0003	0.0000	1.914	0.2799	0.3317	0.1493	0.1163	2.694
0.4	0.3	0.1858	0.2853	0.0002	0.0000	2.024	0.2808	0.3319	0.1502	0.1173	2.702
0.4	0.4	0.1877	0.2542	0.0002	0.0000	1.927	0.2812	0.3320	0.1506	0.1177	2.705
0.4	0.5	0.2160	0.2037	0.0002	0.0000	1.870	0.2814	0.3320	0.1509	0.1180	2.708
0.5	0.0	0.1520	0.2844	0.0003	0.0000	2.001	0.1243	0.2368	0.0284	0.0113	1.858
0.5	0.1	0.1758	0.2832	0.0003	0.0001	2.014	0.2783	0.3310	0.1458	0.1131	2.668
0.5	0.2	0.1872	0.2697	0.0002	0.0000	1.970	0.2803	0.3316	0.1488	0.1160	2.691
0.5	0.3	0.1783	0.2224	0.0002	0.0000	1.875	0.2810	0.3318	0.1498	0.1170	2.700
0.5	0.4	0.2374	0.2068	0.0002	0.0000	1.886	0.2814	0.3319	0.1504	0.1176	2.704
0.5	0.5	<u>0.3046</u>	0.2814	0.0003	0.0000	<u>2.315</u>	0.2816	0.3320	0.1507	0.1179	2.707
1.0	0.0	<u>0.3058</u>	0.2715	0.0001	0.0000	<u>2.291</u>	<u>0.3045</u>	0.1351	0.0033	0.0003	2.168

Publications

Appendix B lists all publications included in this cumulative dissertation and summarizes my scientific and written contributions to each work.

B.1 [P1] Investigating the Impact of Neural Network Regularization for Affect Recognition through Quantified Local Explanations

Ines Rieger (2024). “Investigating the Regularization of Deep Neural Networks for Affect Recognition With Relevance-Guided Local Explanations”. In: *Artificial Intelligence: Methodology, Systems, and Applications: 19th International Conference, AIMSA 2024*. Lecture Notes in Computer Science. Springer, pp. 122–127. DOI: [10.1007/978-3-031-81542-3_10](https://doi.org/10.1007/978-3-031-81542-3_10)

Authorship and origin: This publication is single-authored. The research idea originated from a discussion with Christian Ledig.

My scientific contributions:

- Developed the study design to analyze CorrLoss regularization and its effect on model features using RelEx.
- Implemented the analysis pipeline and conducted all experiments.
- Performed the evaluation and interpreted the results.

My written contributions:

- Wrote and revised the full manuscript.

B.2 [P2] Domain-Specific Evaluation of Visual Explanations for Application-Grounded Facial Expression Recognition

Bettina Finzel, **Ines Rieger**, Simon Kuhn, and Ute Schmid (2023). “Domain-Specific Evaluation of Visual Explanations for Application-Grounded Facial Expression Recognition”. In: *Machine Learning and Knowledge Extraction. CD-MAKE 2023*. Lecture Notes in Computer Science. Springer, pp. 31–44. DOI: [10.1007/978-3-031-40837-3_3](https://doi.org/10.1007/978-3-031-40837-3_3)

Authorship and context: The study was conducted as part of Simon Kuhn’s master’s thesis, co-supervised by Bettina Finzel and myself. The work builds on a research direction introduced as future work in [P5] (**Rieger** et al., [2020a](#)).

My scientific contributions:

- Co-developed the research framing with a focus on domain-specific expert knowledge for evaluating robustness and interpretability. This included the annotation concept by extending bounding boxes to polygonal representations for improved AU localization assessment.
- Implemented and trained the AU detection model (VGG-16 based), including a training strategy for missing labels adapted from Pahl et al. ([2020b](#)). The training data was combined from several datasets, including CK+, Actor Study, Aff-Wild2, BP4D, EmotioNet, and UNBC.
- Provided regular scientific and technical guidance during the thesis work.

My written contributions:

- Wrote Section 3 (materials and methods), excluding the design of Figure 1.
- Co-wrote Section 4 (results and discussion), contributing to the interpretation of the results and their implications.
- I reviewed and refined the manuscript.

B.3 [P3] Female, White, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces

Jaspar Pahl, **Ines Rieger**, Anna Möller, Thomas Wittenberg, and Ute Schmid (2022). “Female, white, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces”. In: *Proc. of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, 973–987. DOI: [10.1145/3531146.3533159](https://doi.org/10.1145/3531146.3533159)

Authorship and context: Jaspar Pahl and I contributed equally to this work, as stated in the publication.

My scientific contributions:

- Co-identified the research gap regarding missing demographic attribute annotations in FER benchmark datasets, enabling systematic fairness evaluation. At the time, testing the fairness of AI models was an emerging research topic (Chen et al., [2024](#)).
- Co-designed the annotation objectives and the evaluation protocol; contributed demographic metadata coding for selected datasets to guide the process.
- Supervised the student assistant Anna Möller during demographic annotation and model evaluation.

My written contributions:

- Reviewed, revised, and provided detailed feedback on the manuscript draft prepared during the project.

B.4 [P4] CorrLoss: Integrating Co-occurrence Domain Knowledge for Affect Recognition

Ines Rieger, Jaspar Pahl, Bettina Finzel, and Ute Schmid (2022). “CorrLoss: Integrating Co-Occurrence Domain Knowledge for Affect Recognition”. In: *Proc. of the 26th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 798–804. DOI: [10.1109/ICPR56361.2022.9956319](https://doi.org/10.1109/ICPR56361.2022.9956319)

My scientific contributions:

- Co-developed the central concept with Jaspar Pahl and Bettina Finzel: improving cross-dataset generalization by integrating AU–AU relationships into the loss function (CorrLoss). This focus on cross-dataset evaluation reflected a growing need in facial expression recognition (Ramis et al., [2022](#)).
- Contributed to the mathematical formulation and implemented CorrLoss in Python.
- Designed and conducted the experimental evaluation (subject-dependent k-fold cross-validation, cross-dataset evaluation, and calibration analyses) across multiple datasets.

My written contributions:

- Served as primary author (approx. 90%): drafted and revised the manuscript, including introduction, related work, method, results, and discussion; created all tables and visualizations.
- Integrated feedback from Jaspar Pahl, Bettina Finzel, and Ute Schmid and finalized the submission.

B.5 [P5] Verifying Deep Learning-Based Decisions for Facial Expression Recognition

Ines Rieger, René Kollmann, Bettina Finzel, Dominik Seuss, and Ute Schmid (2020a). “Verifying Deep Learning-based Decisions for Facial Expression Recognition”. In: *Proc. of the 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pp. 139–144. URL: <https://www.esann.org/sites/default/files/proceedings/2020/ES2020-49.pdf>

My scientific contributions:

- Co-developed the research concept with Bettina Finzel, focusing on quantitative evaluation of AU heatmaps stemming from deep learning models using domain knowledge on AU localization.
- Performed data preprocessing and model training using a modified ResNet-18 model developed in my masters thesis (Rieger et al., 2019).
- Co-supervised the student assistant René Kollmann, who conducted the heatmap evaluations.

My written contributions:

- Wrote Sections 2.1 (datasets and preprocessing) and 2.2 (model description).
- Co-wrote Section 3.1 (results) with René Kollmann and revised the full manuscript.

B.6 [P6] Unique Class Group Based Multi-Label Balancing Optimizer for Action Unit Detection

Ines Rieger, Jaspar Pahl, and Dominik Seuss (2020b). “Unique Class Group Based Multi-Label Balancing Optimizer for Action Unit Detection”. In: *Proc. of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG), Workshop on Affect Recognition in-the-wild: Uni/Multi-Modal Analysis and VA-AU-Expression Challenges*. IEEE, pp. 619–623. DOI: [10.1109/FG47880.2020.00101](https://doi.org/10.1109/FG47880.2020.00101)

Authorship and context: The MLOPT approach achieved third place in the Affective Behavior Analysis in-the-Wild (ABAW) challenge at the IEEE Conference on Face and Gesture Recognition 2020 (Kollias et al., [2020](#)).

My scientific contributions:

- Co-developed the research concept of MLOPT with Jaspar Pahl, targeting multi-label imbalance in AU detection.
- Contributed to model training using a combined AU training dataset and to evaluation on the Aff-Wild2 validation set.
- Co-developed the dataset integration framework STRAND with colleagues to standardize and process heterogeneous annotations across datasets (see Section [3.3](#)). This approach makes it easy to combine datasets while preserving the integrity of the original data, streamlining dataset management for large-scale, multi-dataset projects.

My written contributions:

- Wrote the full manuscript and extended material from the technical challenge paper (Pahl et al., [2020a](#)).
- Created Figure 1.

B.7 [P7] FMC-Net: A Human-Guided Deep Learning Framework for Adaptable and Transparent Facial Expression Recognition in Real-World Scenarios

Ines Rieger, Jaspar Pahl, and Ute Schmid (2025). “FMC-Net: A Human-Guided Deep Learning Framework for Adaptable and Transparent Facial Expression Recognition in Real-World Scenarios”. In: *Applied Intelligence* 55.18. DOI: [10.1007/s10489-025-07017-9](https://doi.org/10.1007/s10489-025-07017-9)

My scientific contributions:

- Co-initiated the core idea with Jaspar Pahl: integrating AU–FE task relationships into training, building on the direction proposed in [P4] (**Rieger et al.**, [2022](#)). Co-developed the mathematical formalization of the AU–FE task-relationship regularization.
- Developed the logic-based component independently, including the rule-based ensemble that combines AU-derived emotion rules with model predictions and yields a disagreement-based risk signal (informed by discussions with Ute Schmid).
- Led the research process, including experimental design, implementation, execution of experiments, and analysis, supported by discussions with Jaspar Pahl and Ute Schmid.

My written contributions:

- Wrote and revised the full manuscript, including figures; incorporated feedback from Jaspar Pahl and Ute Schmid.