Applied Intelligence (2025) 55:1127
https://doi.org/10.1007/s10489-025-07017-9

FMC-Net: A Human-Guided Deep Learning Framework for Adaptable and Transparent Facial Expression Recognition in Real-World Scenarios

Ines Rieger[1,2] · **Jaspar Pahl**[1,2] · **Ute Schmid**[2]

Received: 28 May 2025 / Accepted: 16 November 2025

Abstract

We introduce FMC-Net, a facial expression recognition (FER) framework that leverages the hierarchical relationship between discrete facial muscle movements, known as Action Units (AUs), and Facial Expressions (FEs) by integrating two complementary constraint layers. This framework couples data-driven learning with psychology-grounded structure. First, a training-time correlation constraint aligns the two tasks within a multi-task network by softly regularizing a target statistical relationship. This can improve sample efficiency and generalization, particularly under limited or biased data. Second, an inference-time fuzzy rule layer maps the networks probabilistic AU predictions to FEs using compact, human-editable from psychological research, yielding transparent, per-decision attributions. An ensemble then combines the model and rule-based pathways and exposes a disagreement-based risk score for human-in-the-loop triage. This two-layer constraint integration addresses the limitations of single-mechanism approaches: training-time constraints shape the learned representations but lack case-wise transparency, while inference-time rules explain decisions but cannot improve the underlying features. Experiments across diverse datasets, including in-the-wild video and cross-dataset evaluation, validate our approach. Constraint-guided training consistently produces models that outperform competitive baselines, while the rule-based pathway can provide transparency and actionable risk signals towards reliable deployment. The proposed methodology is also generalizable to other machine learning tasks with interdependent outputs.

Keywords Deep learning · Facial expression recognition · Constraint integration · Human-guided AI

1 Introduction

Facial expression recognition (FER) is a cornerstone of affective computing [1], with applications in medical treatment, driver fatigue monitoring, and psychological research [2]. While deep neural networks (DNNs) have driven performance benchmarks in FER [3–6], their deployment in dynamic, real-world scenarios remains constrained by limited transparency, limited adaptability, and unpredictable behavior under distribution shift. Sensitive applications require AI systems that not only achieve high accuracy, but are also operationally transparent and adaptable to new conditions.

In this work, we adopt the following term definitions: (i) *Transparency* denotes the systems ability to expose a human-readable reasoning path from Action Units (AUs; discrete facial muscle movements defined by the Facial Action Coding System (FACS) [7]) to Facial Expressions (FEs) via an explicit, rule-based inference layer and to output a disagreement-based risk signal. For broader context on transparency and trustworthy AI, see [8–10]. Concretely, transparency in our framework is provided by (a) symbolic, editable AU→FE rules from psychological research that expert users can inspect and modify; (b) per-decision attribution in terms of activated rules and their degrees of support; and (c) risk-aware outputs from an ensemble that quantify prediction confidence. (ii) *Adaptability* denotes the capacity to maintain or improve performance under data

✉ Ines Rieger
ines.rieger@uni-bamberg.de

Jaspar Pahl
jaspar.pahl@iis.fraunhofer.de

Ute Schmid
ute.schmid@uni-bamberg.de

[1] Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

[2] Chair for Cognitive Systems, University of Bamberg, Bamberg, Germany

Published online: 03 December 2025

variability across datasets, domains, and user populations. In FER, adaptability has been explored through alternative mechanisms such as controllable or context-aware adaptation [11, 12]. In our framework, adaptability is achieved by injecting structured AU knowledge during training (regularizing learning under bias or data scarcity) and by enabling user-guided rule adaptation at inference. In experiments, we evaluate adaptability primarily via cross-dataset generalization and low-data robustness and ablate the contributions of constraint-guided training versus rule-based refinement.

FER in-the-wild is characterized by data variability, including subtle, compound, and posed expressions, as well as noise and partial occlusions, yielding models with limited robustness and generalization [13, 14]. Even within a single class, facial expressions vary substantially in appearance and intensity (high intra-class variability). Biased or limited datasets further restrict diversity, degrading cross-dataset reliability [15]. Existing techniques such as temporal or context modeling [13, 14, 16], pseudo-labeling [17], causal modeling [18], and domain constraints [5, 19–21] address specific aspects but often remain black boxes. This leaves open questions of why a prediction was made and how to adapt behavior when conditions shift. This gap is also critical in light of regulatory requirements (e.g., EU AI Act) that emphasize transparency and robustness for emotion recognition [9, 10].

We argue that structured domain knowledge can complement data-driven learning to yield FER systems that are more robust, operationally transparent, and adaptable [9, 22]. A widely accepted representation of facial behavior is the Facial Action Coding System (FACS) [7], which encodes discrete muscle movements as AUs [23–25]. AUs provide a granular, objective substrate underlying basic emotions (e.g., happiness via AU6 cheek raiser and AU12 lip corner puller) as well as compound and dimensional affect, pain, and depression [3, 26–32]. This hierarchical relationship offers a principled avenue to regularize learning and to make inference steps interpretable.

FMC-Net We propose the novel *Facial Movement Constraint Network (FMC-Net)*, a multi-paradigm framework for FER that integrates structured AU knowledge through two complementary technical mechanisms, as illustrated in our conceptual overview (Fig. 1). The framework is instantiated within a multi-task DNN that jointly predicts AUs and FEs:

1. *Constraint-Guided Multi-Task Training* (adaptability). We introduce a hierarchy-aware, inter-task correlation/consistency regularization term that explicitly aligns AU–FE statistics during multi-task training. The target statistics can be defined by users beforehand. This can steer the shared representation to respect known facial dynamics, thereby improving performance and generalization to unseen individuals and domains when training data is limited or biased.
2. *Rule-Based Inference Refinement* (transparency & adaptation). A fuzzy logic module maps probabilistic AU predictions to FE classes using human-readable, editable rules derived from established psychological findings. This provides stepwise, symbolic reasoning that can be inspected and adapted to new populations or contexts with minimal data.

To improve reliability, FMC-Net employs a risk-aware ensemble that fuses the DNNs direct FE predictions with the rule-based FE predictions, producing a disagreement score, and important feature for trustworthy AI in healthcare and social robotics [33, 34].

Because AU annotation and AU–FE mapping require specialized training [7, 25], FMC-Net primarily targets expert users (e.g., clinicians, psychologists) who can define or validate domain-specific constraints, improving both the correctness and the applicability of the AU–FE statistics and rules [35, 36]. At the same time, the framework is designed to minimize expert burden: we use a compact, reusable, soft AU→FE prior derived from EMFACS findings [24] that is reused across datasets without per-dataset rule engineering. The framework is extensible: future applications can incorporate constraints tied to psychological theory, demographic priors, or context-dependent behavior, and can expose simplified interfaces to involve non-expert users where appropriate.

Contributions FMC-Net advances human-guided applied intelligence for FER by unifying data-driven learning with structured, transparent methods:

1. *Operational Transparency:* (a) A rule-based, symbolic AU→FE layer exposes editable reasoning and per-decision rule activations; (b) ensembling communicates disagreement for risk-aware use.
2. *Operational Adaptability:* (a) Constraint-guided training improves performance and low-data robustness by aligning features with AU structure; (b) user-guided rule editing enables rapid adaptation to new domains or populations without full retraining.
3. *Empirical Validation.* We evaluate adaptability via cross-dataset transfer and low-data regimes, quantify transparency through rule-level audits and reporting of risk-aware disagreement, and provide ablations isolating the effect of constraints and rules.
4. *Practicality.* FMC-Nets components are simple to implement and train, facilitating adoption in applied settings.

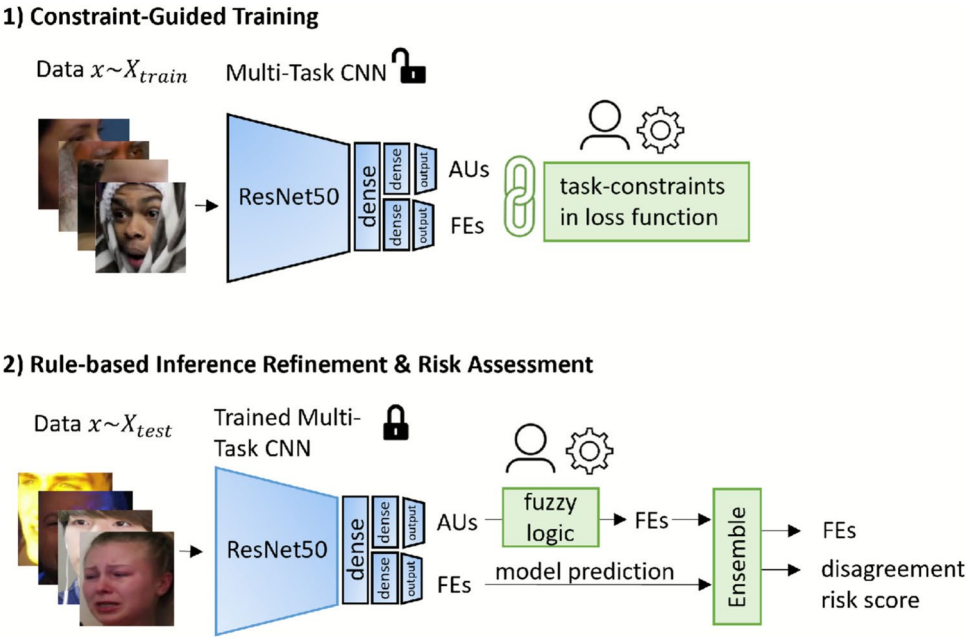

Fig. 1 General framework. Conceptual overview of FMC-Net. 1) During training, a multi-task CNN predicts AUs and FEs while a constraint term encodes AU–FE relationships to regularize shared representations (adaptability). 2) During inference, a fuzzy logic layer maps AU probabilities to FE classes via editable rules (transparency & adaptability). A risk-aware ensemble fuses DNN and rule-based outputs to deliver reliable FE predictions

In summary, FMC-Net demonstrates that multi-layer integration of domain knowledge can improve performance, transparency, and adaptability in complex real-world FER. Although instantiated for FER, the underlying principles generalize to other multi-task problems with structured inter-task relations.

2 Related work

We review strands relevant to FER: human-guided/interactive AI; domain knowledge in deep models and multi-task learning (MTL); correlation- or structure-based constraints; rule-based inference including fuzzy logic; and ensembles with disagreement-based risk. We emphasize representative approaches to situate our work and identify the following recurring limitations across these areas:

1. *Symbol grounding gap:* Human guidance is often collected at the UI or data layer, but not bound to internal reasoning variables (e.g., AUs), so it cannot steer representation learning.
2. *Static vs. adaptive knowledge:* Hard-wired priors (graphs/templates) help in-distribution but are brittle under shift; post-hoc rules alone cannot improve learned features.
3. *Multi-task learning interference:* Sharing without explicit inter-task alignment can cause negative transfer between AU and FE.
4. *Training inference mismatch:* Constrained training improves features but rarely yields per-decision rationales; rules explain decisions but do not shape the representation.
5. *Uncertainty without semantics:* Dispersion scores flag risk, yet are not attributed to AU/FE knowledge, limiting actionability.

2.1 Human-Guided FER and Interactive AI

Human-centered and interactive ML frameworks argue for workflows that keep model behavior inspectable and adaptable for users [37, 38]. In FER, representative interactive systems include Emoticontrol, which adapts application behavior based on inferred user emotions (interface-level adaptation rather than model adaptation) [11], and Emotion AWARE, which lets users customize emotion lexicons for specific applications [12]. Related practices such as interactive labeling or active learning are also discussed in affective-computing surveys [22]. A recurring limitation is symbol grounding: guidance typically occurs at the UI/data layer and is not bound to internal variables such as AUs, making it difficult to steer representation learning or obtain case-wise rationales.

2.2 Single-task FER and temporal modeling

Recent single-task FER baselines include strong image/video backbones; representative examples used in our comparisons are EfficientNet for video FER [39] and semi/self-supervised FER with pseudo-labeling and temporal modeling (SSL+DTM) [17]. Classical temporal architectures that pair 2D CNN encoders with recurrent heads (e.g., ResNet+GRU) remain standard references for sequence modeling and appear in our comparison (ResNet50+GRU as a temporal baseline). For BP4D, a recent single-task FER

reference augments ResNet18 with class-imbalance handling (ResNet18+SMOTE) [40].

2.3 Domain knowledge in multi-task learning and joint AU–FE models

Multi-task learning transfers across related outputs by sharing parameters or selecting auxiliary tasks. Common strategies include adaptive sharing and meta-learning of task relations [20, 41]. Joint AU–FE models use transformer-based encoders shared across tasks or masked autoencoder-style pretraining with multi-task heads to couple representations [4, 42].

Beyond this implicit coupling, prior FER work injects domain knowledge and structure in several ways. Knowledge graphs encode concept relationships and enable supervision to propagate across related nodes (e.g., AUs, FEs, attributes) [43]. Semantic descriptions embedded alongside visual features can tie predictions to human-understandable attributes [44]. Inter-task dependencies can be used as weak supervision to guide learning across outputs [45]. Meta-information as priors biases training toward context-aware or demographically informed behavior [5, 16].

A particularly relevant line of work introduces structured inductive biases via explicit constraints. For facial analysis, causal structures have been used to encourage consistent AU predictions (CISNet) [18]. Our own prior work, CorrLoss [21], demonstrated that a correlation-based constraint could improve generalization by enforcing AU co-occurrence patterns, but this was limited to a single-task AU detection setting. Other multi-task methods have explored meta-learned relations (MAL) [20] or auxiliary modalities like optical flow (JAO) [28] to guide learning.

FMC-Net makes several novel contributions that directly extend and unify these prior concepts. First, it generalizes the single-task correlation constraint from CorrLoss to a more complex, multi-task setting, introducing a novel inter-task (AU-FE) constraint that was not previously explored. Second, unlike methods that fix knowledge at design time or encode it implicitly in shared weights, FMC-Net's entire two-layer architecture is designed for user-adaptability. The training-time constraints and the inference-time rules are both explicit and modifiable. This provides a level of per-decision transparency and post-deployment adaptability that is not present in the other MTL or constraint-based approaches discussed.

2.4 Rule-based inference and fuzzy logic for AU→FE

Psychology-grounded mappings (e.g., EMFACS [24]) specify AU combinations associated with discrete emotions and have long informed computational AU→FE pipelines. Early approaches encoded these mappings as Boolean templates or multi-step schemes [46, 47], and recent FER tools also apply affective appraisals from AU intensities [14].

To interface symbolic knowledge with modern DNN outputs (continuous AU probabilities/intensities), fuzzy logic provides graded memberships and t-norm operators that support soft rule evaluation [48]. Recent neuro-symbolic work has focused on making such logic differentiable and easy to compose with deep features [49]. In applied vision and XAI, fuzzy logic layers have been used to verify concept satisfaction for CNN predictions [50] and to inject symbolic consistency when correcting or filtering noisy labels [51]. Within affective computing specifically, fuzzy rule systems (e.g., ANFIS) continue to appear as interpretable mappers from signals (facial landmarks, text, AU scores) to AUs or emotions, offering user-editable rules with graded confidence [52, 53].

In contrast to works that employ fuzzy logic primarily as a standalone classifier or a post-hoc verifier, FMC-Net integrates a fuzzy AU→FE layer as a complementary inference pathway alongside a novel training-time correlation/consistency constraint. The layers purpose is twofold: (1) provide case-wise transparency via explicit rule activation on continuous AU outputs, and (2) enable lightweight, post-deployment adaptation by editing rules without retraining. Purely rule-driven systems do not reshape learned representations and can be brittle under shift; conversely, training-time constraints alone offer limited per-decision insight. FMC-Net combines both: constraints that shape features during learning and a symbolic pathway that explains and can be adjusted at inference.

2.5 Ensembles and disagreement-based risk

A common way to improve robustness in practice is to aggregate heterogeneous predictors. This averaging tends to smooth errors and reduce variance across seeds or splits, supporting a stability-aware model selection perspective rather than optimizing mean scores alone [54–56]. In FER, ensemble strategies are commonly used to boost performance, though they often report only aggregated gains without analyzing failure modes [57].

Beyond averaging, the disagreement between complementary predictors can be used as a deployable risk signal. Selective prediction frameworks operationalize this idea by accepting a prediction only when some confidence or disagreement is reached [58]. When one of the predictors encodes domain knowledge, the disagreement becomes interpretable. This means, conflicts can be traced to specific rule activations and AU patterns, which provides human-legible cues for triage and debugging.

2.6 Position of FMC-net across strands

Unlike prior work that treats these strands independently, FMC-Net integrates (1) a soft, editable AU↔FE structure into multi-task training, (2) a symbolic AU→FE pathway for transparent inference, and (3) an interpretable reliability cue via model knowledge disagreement. This composition addresses the symbol-grounding, explanation, and actionability gaps in a single, reusable framework, complementing existing multi-task learning, and domain-generalization techniques without requiring per-dataset rule engineering.

3 FMC-Net: Framework architecture and methods

This section details the methodology and architecture of our Facial Movement Constraint Network (FMC-Net), a human-guided AI framework illustrated in Fig. 1. We first describe our methodology for injecting domain knowledge via constraint-guided multi-task training (Sec. 3.1). We then explain how this is complemented by a rule-based fuzzy logic layer for transparent inference (Sec. 3.2) and an ensemble structure that enables disagreement-based risk quantification (Sec. 3.3). Finally, we provide detailed information about the FMC-Net model architecture.

3.1 Constraint-guided multi-task training

FMC-Net integrates standard multi-task learning with user-defined task constraints so that human knowledge can shape the shared representation. This allows human domain knowledge to directly influence the network's learning process, encouraging the model to capture relationships between interrelated tasks that are aligned with expert understanding or data statistics. We consider two interrelated tasks: FER and AU detection. The constraint steers the network to learn inter-task structure (AU↔FE) and, optionally, intra-task AU structure.

Task-specific loss function Let F be the number of FE classes with index $f \in \{1,\dots,F\}$ and A the number of AUs with index $a \in \{1,\dots,A\}$. The FE head outputs probabilities $\hat{y}^{(F)}_{i,f} \in [0,1]$ (softmax; $\sum_f \hat{y}^{(F)}_{i,f} = 1$) with one-hot labels $y^{(F)}_{i,f} \in \{0,1\}$; the AU head outputs $\hat{y}^{(A)}_{i,a} \in [0,1]$ (sigmoid) with binary labels $y^{(A)}_{i,a} \in \{0,1\}$. We use categorical cross-entropy for the FE multi-class classification and binary cross-entropy for the multi-label AU detection, averaged over the batch with size B.:

$$L_F = -\frac{1}{B}\sum_{i=1}^{B}\sum_{f=1}^{F} y^{(F)}_{i,f}\,\log \hat{y}^{(F)}_{i,f}, \tag{1}$$

$$L_A = -\frac{1}{B}\sum_{i=1}^{B}\sum_{a=1}^{A}\Big[y^{(A)}_{i,a}\,\log \hat{y}^{(A)}_{i,a} + (1-y^{(A)}_{i,a})\,\log(1-\hat{y}^{(A)}_{i,a})\Big]. \tag{2}$$

Task-constraint R_{task} To inject structure, we compare predicted batchwise correlations with a target correlation pattern. For this we collect the current batch probabilities as $\widehat{Y}^{(F)} \in \mathbb{R}^{B\times F}$ and $\widehat{Y}^{(A)} \in \mathbb{R}^{B\times A}$. Here, the columns correspond to the classes. We compute correlations column-wise across the sample dimension using either the Pearson Correlation Coefficient (PCC) or the Concordance Correlation Coefficient (CCC), both ranging in $[-1,1]$.

Inter-task (AU↔FE) Let $C^\star \in \mathbb{R}^{A\times F}$ be a target AU–FE correlation matrix, and let $\widehat{C} = \mathrm{corr}(\widehat{Y}^{(A)}, \widehat{Y}^{(F)}) \in \mathbb{R}^{A\times F}$ be the batchwise predicted correlation. Over all unique AU–FE pairs $\Omega_{\text{inter}} = \{(a,f) : 1\le a\le A,\ 1\le f\le F\}$, and therefore $|\Omega_{\text{inter}}| = A\cdot F$:

$$R_{\text{task}} = \left(\frac{1}{|\Omega_{\text{inter}}|}\sum_{(a,f)\in\Omega_{\text{inter}}} |\widehat{C}_{a,f} - C^\star_{a,f}|^p\right)^{1/p}, \qquad p\in\{1,2\}. \tag{3}$$

Intra-task (AU↔AU) Let $C^\star_{AA} \in \mathbb{R}^{A\times A}$ be a target AU–AU correlation matrix and $\widehat{C}_{AA} = \mathrm{corr}(\widehat{Y}^{(A)}, \widehat{Y}^{(A)})$ the batchwise predicted AU–AU correlation. Over off-diagonal unique AU pairs $\Omega_{\text{intra}} = \{(a,b) : 1\le a < b\le A\}$, and therefore $|\Omega_{\text{intra}}| = \frac{A(A-1)}{2}$

$$R_A = \left(\frac{1}{|\Omega_{\text{intra}}|}\sum_{(a,b)\in\Omega_{\text{intra}}} \left|\widehat{C}_{AA,ab} - C^\star_{AA,ab}\right|^p\right)^{1/p}. \tag{4}$$

Intuitively, these constraints shrink the gap between the correlation structure implied by the models current predictions and the desired target correlation. Using CCC instead of PCC is more rigorous and emphasizes agreement in both correlation and location/scale.

Final objective We combine task losses and constraints with a non negative weight α (and a switch $\gamma \in \{0,1\}$ for the AU–AU term):

$$\tilde{L} = (1-\alpha)(L_F + L_A) + \alpha\Big(R_{\text{task}} + \gamma\, R_A\Big). \tag{5}$$

Variable α controls the strength of the constraint regularization and is selected on validation data. Setting $\gamma = 1$ yields the combined constraint (R_{both}), whereas $\gamma = 0$ enforces only AU↔FE structure, called R_{task}. R_{task} and R_A are computed per batch. Minimizing $\tilde{L}$ during training encourages the model to achieve high performance on the AU and FE tasks while simultaneously adhering to the specified constraint, thus promoting the learning of interrelated features.

Single-task AU variant For the single-task AU ablation (Table 7), we use the same intra-AU constraint from (4) and optimize

$$\tilde{L}_{\text{single-AU}} = L_A + \alpha R_A,$$

in the spirit of [21], with $\text{corr} \in \{\text{PCC}, \text{CCC}\}$ and $p \in \{1, 2\}$.

Practical notes (1) Correlations are computed on the predicted probabilities $\widehat{Y}^{(F)}$ and $\widehat{Y}^{(A)}$ (not hard labels), which stabilizes estimates across batches.
(2) Target matrices ($C^\star$ for AU–FE and $C^\star_{AA}$ for AU–AU) can be derived from training-set ground truth, from a representative auxiliary/target dataset, or from user-specified priors; incorporated directly into (3), (4), and (5).
(3) We normalize by the number of unique class combinations (denominators in (3) and (4)) so that penalty magnitudes are comparable across datasets and different F/A, making α more transferable across settings.

3.2 Rule-based facial expression inference via fuzzy logic

FMC-Net includes a transparent inference layer that maps AU predictions to FE classes using a rule set. In all experiments we use EMFACS-derived rules that formalize established psychological links [24] depicted in Table 1. It is possible to use other rule sets for multi-output problems.

Rule set For each FE class f, Table 1 specifies a disjunctive normal form (DNF): a disjunction (OR) of one or more clauses, where each clause is a conjunction (AND) of AUs (e.g., Happiness : $(6 \wedge 12) \vee (7 \wedge 12)$). While our specific rules do not use negations, the implementation supports them.

Fuzzy operators To operate on probabilistic AU outputs $\hat{y}^{(A)} \in [0, 1]^A$, we evaluate rules with Product fuzzy logic [48] (Table 2): AND = product ab, OR = probabilistic sum $a+b-ab$. This yields a raw score $s_f \in [0, 1]$ for each FE class. Fuzzy logic is effective for evaluating conjunctions and disjunctions of DNN outputs [49].

From scores to probabilities Each emotion rule is evaluated independently, producing a raw score. The raw scores are converted to a categorical distribution by ℓ_1 normalization. This produces a rule-based FE probability vector directly comparable to the model heads softmax output $\hat{y}^{(F)}$. The highest resulting class score is selected as FE prediction. The predictions from the two heads can then be averaged for the ensemble or contrasted to calculate the disagreement-based risk (Sec. 3.3).

Boolean baseline For ablations (Sec. 5.1), we include a Boolean variant that thresholds AU activations from the DNN at 0.5 before applying the logical AND/OR.

Implementation details Rules: Table 1 (no learned weights). Fuzzy operators: Product t-norm / probabilistic t-conorm; no temperature or sharpening; ℓ_1 normalization only. Numerics: Clip probabilities to $[10^{-6}, 1 - 10^{-6}]$ before products.

Transparency and adaptability By integrating fuzzy logic, this layer provides a transparent method for inference. The rule structure is explicit, and the fuzzy logic evaluation process is clear. Furthermore, it enables user-adaptability at the inference stage. The transparency of the rule structure and the ability for real-time modification are particularly valuable for identifying and mitigating biases that may be present in the data-driven models predictions or inherent in fixed mapping rules. This can improve performance or alignment with specific

Table 1 Logical rules to map AUs to FEs from EMFACS [24]

Emotions	Action Units
Anger	$(4 \wedge 7) \vee (4 \wedge 5) \vee (4 \wedge 5 \wedge 7) \vee (17 \wedge 24)$
Fear	$1 \wedge 2 \wedge 4$
Disgust	$9 \wedge 15 \wedge 16$
Happiness	$(6 \wedge 12) \vee (7 \wedge 12)$
Sadness	$(1 \wedge 4) \vee (6 \wedge 15) \vee (11 \wedge 15) \vee (11 \wedge 17)$
Surprise	$(1 \wedge 2 \wedge 5) \vee (1 \wedge 2 \wedge 26)$

Table 2 Product Fuzzy Logic interpretation of logical connectives [48]

Logical	Connective	Interpretation	
$\alpha \wedge \beta$	conjunction (AND)	$\alpha \cdot \beta$	t-norm
$\alpha \vee \beta$	disjunction (OR)	$\alpha + \beta - \alpha \cdot \beta$	t-conorm

application requirements in interactive HCI systems. This user-adaptive rule integration forms the second layer in FMC-Net's multi-layer constraint design, complementing the training-level constraints.

3.3 Ensemble prediction and disagreement-based risk score

A key component of FMC-Net for improving robustness and transparency in applied settings is its ensemble prediction method and the associated disagreement between the models. We compare two heads for FE prediction: a data-driven model head and a knowledge-driven rule-based head (fuzzy logic over AUs). Let C be the number of FE classes and, for each instance i, let

$$p_{\text{model},i} \in \Delta^{C-1}, \qquad p_{\text{rule},i} \in \Delta^{C-1},$$

denote the corresponding class-probability vectors. Rule-based outputs are ℓ_1-normalized so that $\sum_c p_{\text{rule},i}(c) = 1$, making them comparable to $p_{\text{model},i}$.

Ensemble prediction We combine the two FE distributions by a convex mixture

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

We fix $\eta = \frac{1}{2}$ for simplicity. A per-dataset η tuned on validation is straightforward but not investigated here.

Disagreement-based risk We quantify model rule disagreement d per class as

$$d_i(c) \;=\; \left| p_{\text{model},i}(c) - p_{\text{rule},i}(c) \right| \in [0,1],$$

which provides a class-specific risk indicator. Higher values indicate stronger divergence between the learned head and the rule set, serving as a warning signal to human users in real-world, potentially high-risk scenarios. We refer to these quantities as disagreement-based risk scores to emphasize that they capture semantic disagreement rather than a decomposition of uncertainty.

Confidence baseline We compare our disagreement-based risk to the normalized entropy of the model head,

$$\hat{H}_i \;=\; -\frac{1}{\log C} \sum_{c=1}^{C} p_{\text{model},i}(c)\, \log p_{\text{model},i}(c) \;\in [0,1],$$

where larger $\hat{H}_i$ indicates a more diffuse (less confident) distribution.

Computational cost Unlike deep ensembles that require training multiple networks, this approach uses a single trained model plus the rule set; disagreement is computed at inference time with negligible overhead.

3.4 Network architecture and inference pipeline

While recent FER approaches increasingly adopt transformer-based architectures due to their ability to model long-range dependencies [4, 57], our framework is designed to be model-agnostic and can, in principle, be integrated with various backbone architectures. Figure 2 illustrates the architecture. For this work, we adopt a convolutional neural network (CNN) architecture for its practical advantages in computer vision tasks, including strong inductive biases for spatially local features and computational efficiency, which are compatible with our interpretable, constraint-guided components. We implement it using a ResNet50 [59] backbone, pretrained on ImageNet [60]. We remove the standard top layers and use the output of the final average pooling layer as the shared feature vector.

On top of this backbone, we attach a custom multi-task head, which is detailed in Fig. 2. The ResNet backbone feeds into a shared dense layer with 256 neurons and a ReLU activation function. This is followed by two parallel, task-specific heads for AU detection and FE recognition. Each of the two heads consists of its own dense layer (256 neurons, ReLU activation) and a final output layer. For regularization, a dropout layer [61] with a probability of 0.5 is applied after each of the three dense layers (the shared one and both task-specific ones). The AU head uses a sigmoid activation function for multi-label binary classification, while the FE head uses a softmax activation function for multi-class classification. All new dense layers are initialized using the Glorot uniform initializer (also known as Xavier initialization) [62] to promote stable training. The standard task losses for the FE and AU branches, L_F and L_A, are summed with equal weighting in our final objective function, as shown in (5).

During inference, this network produces two distinct outputs. The FE head provides a data-driven probability vector (p_{model}). In parallel, the vector of probabilistic AU predictions from the AU head serves as the input to our transparent, rule-based fuzzy logic layer (detailed in Sec. 3.2), which translates these AU predictions into a knowledge-driven FE probability vector (p_{rule}). Finally, as described in Sec. 3.3, these two probability vectors are combined in an ensemble to produce the final prediction and to calculate a disagreement-based risk score.

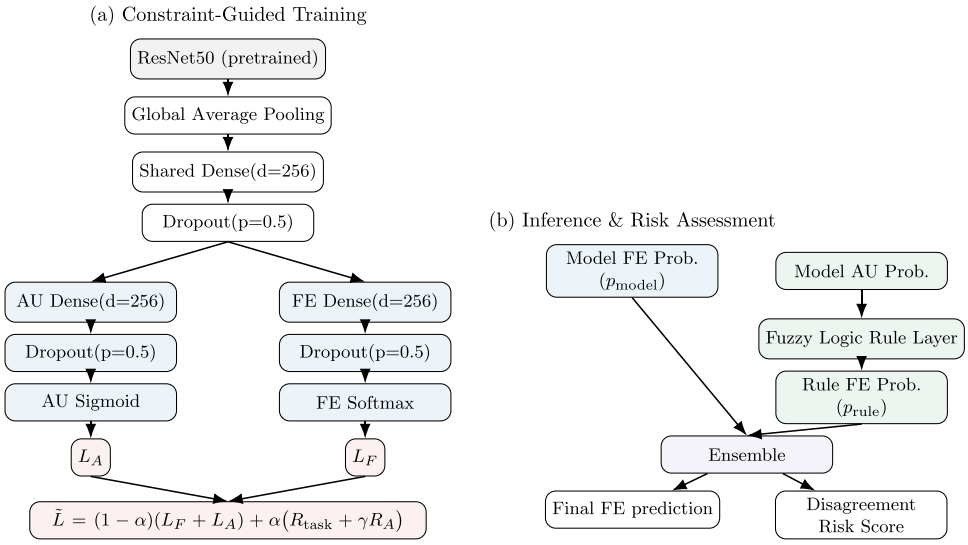

Fig. 2 Schematic of the FMC-Net framework and its two operational layers. **(a) Constraint-Guided Training:** A multi-task network with a shared head and task-specific branches is trained with a composite loss function, $\tilde{L}$, that combines standard task losses (L_A, L_F) with our proposed correlation-based constraint terms (R_{task}, R_A). **(b) Inference & Risk Assessment:** The trained network's two outputs (p_{model} and AU probabilities) are used by the fuzzy logic layer and ensemble to produce the final prediction and a disagreement-based risk score

4 Experimental setup

This section details the experimental setup used to evaluate FMC-Net and baseline models. We present the datasets utilized, the preprocessing steps applied, the training configuration, and the evaluation metrics employed. Together, this information provides a reproducible framework.

4.1 Datasets and preprocessing

We employ four public benchmark datasets with diverse characteristics to ensure a comprehensive evaluation for k-fold cross-validation and cross-dataset evaluation. An overview is provided in Table 3, and the final class distributions used in our experiments are detailed in Tables 4 and 5.

Dataset characteristics (Table 3) We use two controlled, laboratory-based datasets for in-depth analysis: the *Actor Study Database* [63], which features 21 subjects with a mix of posed and spontaneous expressions, and the larger *BP4D-Spontaneous Dataset* [64], which contains 41 subjects exhibiting purely spontaneous emotions. To test generalization to real-world conditions, we use the *Aff-Wild2 Database* [65], a large, challenging dataset captured entirely *in-the-wild* from online videos. Finally, the small, lab-based *extended Cohn-Kanade (CK+)* dataset [66] is used solely as a cross-dataset test target.

Data selection and denoising (Tables 4 and 5) To ensure methodological consistency and address known data quality issues, we applied the following selection protocol. For k-fold cross-validation, we used only frames annotated with both FEs and AUs, selecting classes with a minimum frame count to ensure sufficient data per class for model training (3,000 for Actor Study and Aff-Wild2; 7,000 for BP4D). Furthermore, to mitigate potential label noise from the sequence-level annotations in Actor Study and BP4D, we selected only frames where at least two FE-related AUs were active. This automated denoising step, designed to increase the likelihood that the frame corresponds to the sequence-level label, was not required for the frame-annotated Aff-Wild2 data. In contrast, other approaches, such as [45], address this same limitation by manually selecting apex (highest activation) frames, a process which resulted

Table 3 Overview of datasets used for k-fold cross-validation and cross-dataset evaluation

Dataset	Type	Size	Labels	Features	Usage
Actor Study [63]	posed/ spont.	777 videos, 21 subj	seq FE & AU	laboratory, age & gender diversity	k-fold CV (FE+AU)
BP4D [64]	spont	328 videos, 41 subj	seq FE, AU	laboratory, narrow age (20-30), ethnic diversity	k-fold CV (FE+AU)
Aff-Wild2 [65]	natural	large & diverse source	frame FE, AU, V&A	in-the-wild, YouTube videos, diversity (age, ethnic, pose, illumination), partial occlusion	k-fold CV (FE+AU subset), cross-dataset test (FE)
CK+ [66]	posed/ natural	123 subjects	frame basic emotion & AU	laboratory, small size	cross-dataset test (FE)

Abbreviations: AU (Action Units), CV (Cross-Validation), FE (Facial Expressions), Seq (Sequence), Spont (Spontaneous/Elicited), V/A (Valence/Arousal)

Table 4 Detailed dataset distribution for Action Units (AU)

Dataset	Split	AU1	AU2	AU4	AU6	AU7	AU12	AU15
Actor Study	k-fold	7,085	4,595	7,333	4,828	6,378	4,650	-
Aff-Wild2	k-fold	10,952	-	5,603	3,406	-	7,146	-
BP4D	k-fold	14,323	10,300	9,157	25,955	30,260	30,423	8,509

This table presents the number of frames per selected AU class used for k-fold cross-validation (k-fold) and testing (test). Empty cells indicate the class was not selected for this split/dataset based on our criteria (e.g., insufficient frames annotated with both FE and AU for k-fold, or not annotated at all for test)

in a much smaller subset of 803 frames for BP4D in their work.

All input images were resized to 224x224 pixels, with faces extracted using the OpenCV DNN face detection module [67].

4.2 Training and evaluation protocol

For training, we do not freeze any layers of the pretrained ResNet backbone, allowing the entire network to adapt and learn new patterns specific to facial expression data. We use the AMSGrad optimizer [68] for its convergence properties. We conduct training over a fixed maximum of 50 epochs for each model configuration, using early stopping (with a patience of 10 epochs) to halt training once the model performance on the validation set stops improving, thereby maximizing the model's potential performance while preventing overfitting. Other core hyperparameters for each model (learning rate and batch size) are determined through a grid search that evaluates all possible combinations of batch sizes $[64, 128]$ and learning rates (LR) $[10^{-3}, 10^{-4}, 10^{-5}]$. The optimal learning rate found across all models and datasets is 10^{-4}. The optimal batch size varies slightly, identified as 64 for the single-task models for Aff-Wild2 and the single-task AU model for BP4D, and 128 for all other models. This optimal configuration for LR and batch size is then used as the basis to identify the best regularization weight α for the constraint-guided models, drawn from the range $[0, 1]$ with a step size of 0.1. The optimal hyperparameter configuration for α found for each model variant and dataset split is presented in Table 6. By individually optimizing these hyperparameters for each model setup (single-task, multi-task, different constraints) and dataset, we ensure a fair comparison of the different approaches. All experiments are run on a single NVIDIA RTX A5000 GPU with fixed random seeds for reproducibility.

Our primary evaluation metric for both the AU and FE tasks is the (macro-averaged) F1-score, chosen for its robustness to potential class imbalance in the datasets. For the multi-class FER task, we also report the overall categorical accuracy and the macro-averaged Area Under the Receiver Operating Characteristics (AUROC) curve, calculated using a One-vs-Rest approach. For all metrics, higher values indicate better performance. Probabilistic outputs from the network are converted to class labels using a standard 0.5 threshold for AU detection and the highest probability score (argmax) for FE recognition.

To account for the high variability in facial expressions across individual subjects and to ensure a reliable evaluation of the model's generalization capability to unseen individuals, we employ a k-fold cross-validation strategy with subject-dependent splits, where all data from one subject is contained within a split. This prevents subject leakage between training and validation/test sets. For the Actor Study and BP4D datasets, we use $k = 3$ folds. For the Aff-Wild2 dataset, we use $k = 2$ folds due to the nature of its in-the-wild video structures, which would otherwise lead to highly uneven subject distributions across folds. All splits strictly partition subjects such that no subject appears in more than one fold. This challenging but realistic evaluation protocol

Table 5 Detailed dataset distribution for Facial Expressions (FE)

Dataset	Split	Anger	Fear	Happiness	Sadness
Actor Study	k-fold	1,779	3,216	4,224	3,486
Aff-Wild2	k-fold	-	-	8,377	14,731
Aff-Wild2	test	7,786	9,023	55,640	35,965
BP4D	k-fold	-	17,477	15,816	7,240
CK+	test	1,022	546	1,331	547

This table presents the number of frames per selected emotion class used for k-fold cross-validation (k-fold) and testing (test). Empty cells indicate the class was not selected. Note that for Aff-Wild2 test, a broader range of FE-annotated frames is used compared to the k-fold splits which require simultaneous AU annotations

Table 6 Optimal hyperparameter configuration for regularization weight α

Single-Task AU Model				
Dataset	PCC(L1)	PCC(L2)	CCC(L2)	
Actor Study	[0.3,0.1,0.6]	[0.3,0.6,0.4]	[0.1,0.6,0.1]	
Aff-Wild2	[0.6,0.5]	[0.7,0.2]	[0.9,0.7]	
BP4D	[0.1,0.2,0.1]	[0.7,0.1,0.4]	[0.1,0.1,0.4]	
Multi-Task Model				
Dataset	PCC		CCC	
	R_{task}	R_{both}	R_{task}	R_{both}
Actor Study	[0.9,0.3,0.7]	[0.4,0.2,0.1]	[0.6,0.8,0.6]	[0.4,0.8,0.1]
Aff-Wild2	[0.5,0.6]	[0.3,0.5]	[0.5,0.3]	[0.3,0.5]
BP4D	[0.3,0.4,0.5]	[0.7,0.6,0.3]	[0.2,0.5,0.2]	[0.2,0.8,0.4]

Values are listed for each dataset and constraint type. For k-fold cross-validation, multiple values are listed corresponding to the optimal α found for each fold

is expected to produce higher variance across folds (i.e., a larger standard deviation) compared to subject-dependent evaluations, reflecting the natural variability in expressions across individuals. All reported metric values are the average across the k-folds.

5 Results

This section systematically evaluates the key contributions of FMC-Net. Our analysis begins with the core contribution of this work: a detailed examination of our constraint-guided training mechanism across diverse datasets, demonstrating its effectiveness in improving multi-task learning for FER. We then provide detailed ablation studies to justify our architectural and methodological choices, including the use of fuzzy logic for the rule-based inference layer. Building on these findings, we evaluate the performance of the fuzzy logic component and analyze the framework's risk scoring capabilities. Finally, we situate FMC-Net's performance by benchmarking it against competitive methods in both within- and cross-dataset scenarios.

5.1 Ablation study: Constraint variants and rule-based logic

Table 7 reports an ablation study evaluating variants of the R_A regularization for single-task AU detection and comparing Boolean versus fuzzy logic for rule-based emotion prediction. The goal is to select the optimal loss variant and rule interpretation for the full multi-task FMC-Net.

Left (AUs) We applied constraints to single-task AU models to enforce relationships between AU classes, testing p-norms (L1, L2) and correlation choices (PCC, CCC). The results consistently show that the L2 norm yields the best AU macro F1 scores. On Actor Study, both the PCC and CCC variants achieved a top mean performance of 79.4, with the L2-PCC model showing high stability (*), indicating the smallest SD among methods within 0.5 points of the best mean in that column for the respective dataset. On BP4D, the L2-CCC model provided both the highest mean and stability (*). Since L2-PCC is also stronger on Aff-Wild2, we adopt the L2 norm for all subsequent constraints and carry both PCC and CCC forward for further evaluation.

Right (Rule-based emotion recognition) We compare a Boolean interpretation of AU-FE rules (thresholding AUs at 0.5) with a fuzzy logic interpretation that uses the continuous AU probabilities (see Section 3.2). Rules are evaluated using either ground-truth AUs (GT) or predicted AUs. This evaluation is shown for BP4D and Actor Study, as the Aff-Wild2 subset has insufficient AU annotations for meaningful rule-based evaluation. The fuzzy logic interpretation consistently and substantially outperforms the Boolean approach when using predicted AUs. Moreover, it surpasses the accuracy of Boolean rules computed with perfect ground-truth AU labels (e.g., on Actor Study, fuzzy logic reaches 60.2 vs. the GT-Boolean 43.7; on BP4D, it reaches 52.0 vs. 33.5). The asterisk markings highlight that these top-performing fuzzy logic models are also the most stable in their respective categories. Given its superior performance and stability, we use the fuzzy interpretation for all subsequent rule-based inference in FMC-Net.

Table 7 Ablation study of single-task AU networks and rule-based emotion predictions with diverse constraint methods

Regularization			AUs	Rule-based Emotion Recognition	
				Boolean	Fuzzy Logic
Method	Corr.	Norm	macro F1	Acc.	Acc.
Actor Study Database					
GT	-	-	-	43.7	-
None	-	-	78.4 ± 3.9	**44.4 ± 6.2***	58.8 ± 1.7
R_A	PCC	L1	78.7 ± 3.7	41.5 ± 6.0	57.5 ± 5.4
R_A	PCC	L2	**79.4 ± 2.6***	42.7 ± 5.9	**60.2 ± 0.8***
R_A	CCC	L2	**79.4 ± 2.9**	43.6 ± 4.9	59.7 ± 3.4
Aff-Wild2 Database					
GT	-	-	-	-	-
None	-	-	37.9 ± 5.1	-	-
R_A	PCC	L1	45.7 ± 0.2	-	-
R_A	PCC	L2	**46.5 ± 0.4***	-	-
R_A	CCC	L2	40.2 ± 3.6	-	-
BP4D-Spontaneous Dataset					
GT	-	-	-	33.5	-
None	-	-	67.0 ± 2.1	**35.1 ± 4.2***	49.6 ± 3.5
R_A	PCC	L1	66.4 ± 1.0	30.2 ± 6.9	**52.0 ± 2.3***
R_A	PCC	L2	68.3 ± 1.8	25.1 ± 4.7	46.6 ± 5.3
R_A	CCC	L2	**68.5 ± 0.6***	28.7 ± 5.8	50.9 ± 2.2

Values are mean ± SD across k folds. **Bold** = highest mean. Asterisk (∗) = smallest SD among methods within 0.5 points of the best mean in that column (stability among top performers)

5.2 Effectiveness of task-constrained training

These experiments demonstrate the impact of our proposed task-constrained multi-task training on FER performance. We evaluate our methods on a variety of datasets with distinct characteristics, detailed in Table 3, to demonstrate the adaptability and effectiveness of the proposed mechanisms. It is important to contextualize the results in Table 8 within this experimental design. All evaluations use a subject-independent k-fold cross-validation setup to test

for generalization to unseen individuals. This challenging protocol, combined with the varying nature of the datasets, influences the performance variance. Specifically, the Actor Study dataset has the smallest subject pool (N=21), which can lead to high variance between folds. The Aff-Wild2 dataset is captured *in-the-wild*, introducing significant variability from uncontrolled head poses, lighting, and occlusions. In contrast, the BP4D dataset, with its larger subject pool (N=41) and controlled laboratory setting, often results in more stable performance metrics.

Table 8 provides a comprehensive comparison of single-task models (ST) versus multi-task models (MT) with and without the proposed training constraints ($R_{\text{task}}, R_{\text{both}}$). R_{both} includes both R_{task} and R_A. We analyze both PCC and CCC correlation types with the L2 norm, as determined in Section 5.1. A combined metric, $(AU_{F1} + FE_{F1})/2$, facilitates a quick assessment of overall performance. The asterisk (*) indicates the smallest standard deviation (SD) among methods within 0.5 points of the best mean in that column for the respective dataset. For a comparison with relevant competitive approaches, please refer to Section 5.5.

5.2.1 Overall findings: Constraints boost multi-task performance

The results in Table 8 consistently show that our proposed constraints are effective. Across all three datasets, the models achieving the best mean performance (marked in bold) are multi-task networks enhanced with our regularization methods. These constrained models not only mitigate the performance degradation sometimes seen in unconstrained multi-task learning but also consistently outperform

Table 8 Comparison of task correlation constraints

	Reg.		Action Units (AU)			Facial Expressions (FE)					Combined Score
Model	Method	Corr.	F1	Prec.	Rec.	F1	Prec.	Rec.	Acc.	AUROC[a]	Best Model
Actor Study Database											
ST	None	-	**78.4 ± 3.9***	79.8 ± 4.5	**77.7 ± 5.0***	73.2 ± 1.7	76.7 ± 1.6	72.5 ± 2.7	75.3 ± 1.2	0.92	75.8 ± 2.4
MT	None	-	76.5 ± 5.3	78.4 ± 6.9	75.9 ± 4.9	72.1 ± 5.4	74.8 ± 4.1	72.1 ± 5.9	74.1 ± 3.8	0.93	74.3 ± 5.2
MT	R_{task}	PCC	**78.4 ± 6.3**	**81.4 ± 8.3**	76.2 ± 5.1	76.7 ± 7.5	81.3 ± 6.4*	74.4 ± 8.2	**79.9 ± 6.1**	0.94	**77.6 ± 6.4***
MT	R_{task}	CCC	77.6 ± 5.7	80.1 ± 6.3	76.3 ± 6.4	75.3 ± 5.6	80.8 ± 3.9	74.2 ± 8.0	78.2 ± 2.6	**0.95**	76.5 ± 5.3
MT	R_{both}	PCC	75.9 ± 5.6	78.0 ± 7.3	74.9 ± 4.8	**76.9 ± 7.6**	**81.7 ± 6.4***	75.0 ± 8.1	**79.9 ± 5.4***	0.93	76.4 ± 5.6
MT	R_{both}	CCC	77.3 ± 2.5	**81.4 ± 5.3***	75.0 ± 3.4	76.6 ± 5.3	78.8 ± 5.3	**76.1 ± 5.9***	79.3 ± 4.8	0.94	77.0 ± 3.4
Aff-Wild2 Database											
ST	None	-	37.9 ± 5.1	47.6 ± 9.5	33.4 ± 2.6	38.8 ± 5.9	44.4 ± 0.3	45.9 ± 1.1	42.4 ± 6.4	0.44	38.4 ± 6.7
MT	None	-	39.1 ± 10.2	**50.6 ± 4.3***	42.0 ± 16.5	43.9 ± 0.2	51.8 ± 6.3	51.2 ± 4.2	50.0 ± 4.9	0.45	41.5 ± 5.2
MT	R_{task}	PCC	**40.6 ± 4.6***	28.6 ± 3.3	**80.5 ± 15.3***	55.0 ± 5.2	66.3 ± 2.6	57.4 ± 3.0	67.0 ± 0.7	0.50	47.8 ± 4.9
MT	R_{task}	CCC	35.9 ± 8.0	39.0 ± 8.4	46.4 ± 20.4	47.7 ± 9.4	44.7 ± 13.7	53.5 ± 3.5	63.0 ± 0.9	0.42	41.8 ± 8.7
MT	R_{both}	PCC	35.7 ± 15.0	37.4 ± 13.3	54.0 ± 20.7	**71.6 ± 1.1***	**71.5 ± 1.2***	**71.9 ± 0.7***	**73.4 ± 1.9***	**0.76**	**53.7 ± 8.0***
MT	R_{both}	CCC	35.6 ± 4.8	43.4 ± 3.1	45.7 ± 18.1	55.4 ± 0.9	58.3 ± 1.9	57.9 ± 2.8	58.7 ± 2.3	0.55	45.5 ± 2.9
BP4D-Spontaneous Dataset											
ST	None	-	67.0 ± 2.1	68.0 ± 5.2	**72.0 ± 5.2***	69.9 ± 3.7	72.1 ± 6.5	69.5 ± 2.5	68.4 ± 4.5	0.85	68.5 ± 2.7
MT	None	-	60.3 ± 4.0	66.4 ± 3.6	60.1 ± 5.0	71.5 ± 1.2	72.5 ± 2.8	71.8 ± 0.6	69.7 ± 1.3	0.85	65.9 ± 1.8
MT	R_{task}	PCC	65.0 ± 4.3	66.7 ± 3.4	65.7 ± 5.3	**73.2 ± 1.6***	73.5 ± 1.8	**73.3 ± 1.7***	**70.8 ± 2.5***	0.85	69.1 ± 2.1
MT	R_{task}	CCC	66.2 ± 1.6	68.6 ± 3.1	66.6 ± 4.0	72.7 ± 4.1	74.3 ± 4.8	73.0 ± 2.7	70.5 ± 4.2	**0.86**	**69.5 ± 1.7***
MT	R_{both}	PCC	**68.0 ± 2.8***	68.1 ± 5.3	71.2 ± 0.5	69.1 ± 3.0	72.0 ± 3.0	68.8 ± 3.4	67.4 ± 1.8	**0.86**	68.6 ± 2.7
MT	R_{both}	CCC	65.3 ± 1.8	**69.6 ± 3.3***	63.9 ± 1.3	71.4 ± 1.3	**74.5 ± 2.8***	71.4 ± 1.3	69.1 ± 0.5	0.85	68.4 ± 1.3

Values are mean ± standard deviation (SD) unless noted. **Bold** indicates the highest mean in each column. An asterix (*) marks the smallest SD among methods whose mean lies within 0.5 points of the best mean in that column (stability among top performers). Because we have ≤ 3 runs, we do not report formal significance tests. AUROC lacks run-level SD; we therefore bold only the maximum. Regularization uses the L2 norm. The Best Model Score is computed per run as $(AU_{F1} + FE_{F1})/2$ and then averaged. Abbreviations: ST (Single-Task), MT (Multi-Task).
[a] AUROC standard deviation is not reported due to an implementation constraint

independently trained single-task models on the combined score. While PCC-based constraints often achieve the highest peak performance, CCC-based constraints are highly competitive and can deliver the best results on other datasets (e.g., BP4D), highlighting the value of multiple correlation measures.

5.2.2 Analysis of performance versus stability

Including standard deviations (SDs) reveals the trade-off between peak performance and stability. On BP4D, the constrained MT R_{task} (CCC) model achieves both the highest Best Model Score and the smallest SD among close contenders ($69.5 \pm 1.7^{*}$ vs. 69.1 ± 2.1), indicating that peak performance and stability coincide.

On Actor Study, by contrast, the MT R_{task} (PCC) model attains the highest mean ($77.6 \pm 6.4^{*}$), while MT R_{both} (CCC) yields a slightly lower mean (77.0 ± 3.4) with markedly lower variability. Although 77.0 lies just outside the 0.5-point proximity used for the stability marker, the comparison illustrates that CCC-based models can be more stable even when they do not achieve the very top mean.

5.2.3 Facial expression recognition gains–strongest on in-the-wild data

A consistent finding across datasets is that the FE task benefits most from our proposed constraints. As shown in Table 8, the top-performing regularized MT models achieve substantial gains in all FE metrics over both the ST and unconstrained MT baselines. This supports the hypothesis that using the anatomically-grounded AU task as a source of structured knowledge can improve the learned representations for the primary goal of expression recognition.

This effect is most pronounced on the challenging Aff-Wild2 dataset. Here, the MT $R_{\text{both}}PCC$ model elevates the FE F1-score from a baseline of 38.8 (ST) to 71.6 ± 1.1. This result is not only the highest-performing but also highly stable. It is important to note that the k-fold validation task for Aff-Wild2 was defined on a smaller set of classes (two emotions, four AUs) due to annotation availability. The magnitude of this performance gain is therefore likely attributable to both the power of our combined constraint method (R_{both}) and the reduced complexity of the classification task itself. Nonetheless, the result provides evidence that our approach can be effective for learning in challenging, in-the-wild conditions.

5.2.4 Confusion matrix analysis on BP4D

Figure 3 provides a qualitative, class-level view for BP4D. The baseline single-task (ST) network (left) shows reasonable performance. The unconstrained multi-task (MT) network (middle) illustrates a common trade-off: its overall FE F1 improves slightly over ST ($69.9 \rightarrow 71.5$), but this comes with weaker AU metrics and a lower combined score (Table 8). This manifests as fewer correct predictions for fear.

Introducing our constraint addresses this trade-off. The constrained MT R_{task} (CCC/L2) model (right), which achieves the highest Best Model Score on BP4D, shows more correct predictions for fear and sadness than both ST and unconstrained MT, and reduces confusions off the main diagonal (e.g., fear→happy). This qualitative pattern is consistent with the quantitative results, where this model attains the top combined score (69.5 ± 1.7*). Overall, the constraint-guided training steers the model toward a more robust and accurate solution on BP4D.

5.2.5 Computational efficiency

Our multi-task model is also efficient compared to deploying separate single-task models. The FER single-task model has 24,113,284 parameters, while the multi-task counterpart uses 24,246,410. This modest increase is outweighed by solving two tasks in a single network, yielding a more resource-efficient solution without substantial computational overhead.

5.3 Within dataset: Rule-based emotion prediction using fuzzy logic

This section evaluates whether constraint-guided training improves the AU inputs used by the fuzzy rule layer. We feed the same fuzzy rules with AUs from two sources: (1) an unconstrained multi-task baseline (MT) and (2) the constrained FMC-Net configured with the best AU constraints per dataset from Table 8.

Table 9 shows that the constrained model is competitive with, and directionally better than, the unconstrained baseline. It has small, consistent gains in both accuracy and F1 (Actor Study: +0.6 Acc / +0.4 F1; BP4D: +1.2 Acc / +1.1 F1). This pattern is in line with Table 8, where the most pronounced benefits of the constraints were observed on the FE task, suggesting that better-structured AU predictions translate into slightly stronger rule-based FE outcomes.

Figure 4 gives a qualitative analysis of the rule-based layers behavior. While performance is strong for emotions like happy and sad, the layer underperforms on fear. We hypothesize that the fear rule ($1 \wedge 2 \wedge 4$) is comparatively strict: as the only rule requiring three AUs jointly with a logical AND, small prediction errors on any constituent AU can zero out the rule output. This discrepancy illustrates a benefit of our human-guided approach: it makes the model's

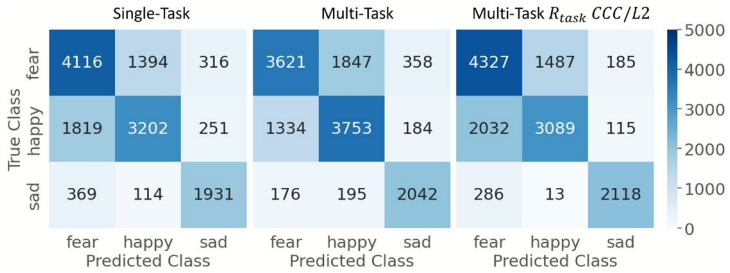

Fig. 3 Confusion matrices on the BP4D dataset comparing three key models from Table 8: the Single-Task baseline (left), an unconstrained Multi-Task model (middle), and our constraint-guided MT model (R_{task} CCC, right). The progression visually demonstrates how our proposed constraint resolves the performance trade-offs of standard multi-task learning, notably improving classification for challenging emotions like fear. Matrices are averaged across all folds

reasoning transparent in this layer, revealing a specific, interpretable weakness in the encoded knowledge (the fear rule) and thus providing a clear target for an expert to refine and improve the system.

5.4 Cross-dataset: Ensemble prediction and disagreement-based risk

This section presents a cross-dataset evaluation to test the generalization of our FMC-Net framework under domain shift. Models trained on controlled laboratory datasets (Actor Study, BP4D) are tested on unseen target domains: the challenging *in-the-wild* Aff-Wild2 validation set and the posed/non-posed CK+ dataset. See Table 5 for the test dataset distributions.

5.4.1 Ensemble prediction

Table 10 compares three distinct prediction methods derived from our best constrained FMC-Nets (Table 8): (1) the direct data-driven model-based head, (2) the transparent knowledge-driven rule-based head, and (3) a our final ensemble that averages the two heads.

Generalization from actor study When transferring from the Actor Study dataset, the results in Table 10 highlight the complementary nature of the different predictors. No single method is universally superior. Notably, the rule-based head demonstrates strong generalization for certain classes, achieving the highest mean F1-score for anger on the Aff-Wild2 and, perhaps more surprising, for the subtle emotion of sadness on both target datasets. However, its F1-score of 0.0 for fear indicates an inability to classify this emotion under domain shift, revealing the sensitivity of its specific logic rule. In contrast, the data-driven model-based head performs well on fear and happy. The ensemble provides a robust balance, achieving the highest overall average score on CK+ (38.9) with the highest stability (*), and outperforming the ResNet50+GRU baseline from [69], a multi-task approach using a similar backbone but trained on a different combined dataset. Similarly on Aff-Wild2, the mean of the ensemble is highly competitive with the best mean and the model is the most stable (*).

Table 9 Within-dataset evaluation of fuzzy rule-based FE predictions using predicted AUs from two sources: an unconstrained multi-task baseline (MT) and the constrained FMC-Net

Model (AU source)	Acc.	F1
Actor Study Database		
MT (unconstrained)	58.3 ± 4.1	49.3 ± 3.7
FMC-Net (constrained)	**58.9** ± **3.9**	**49.7** ± **4.1**
BP4D-Spontaneous Dataset		
MT (unconstrained)	51.1 ± 0.9	47.5 ± 0.9
FMC-Net (constrained)	**52.3** ± **1.1**	**48.6** ± **1.2**

Values are mean ± SD across *k* folds. Best means are in **bold**

Generalization from BP4D Training on the larger BP4D dataset reveals the power of the rule-based system in a different light. On the challenging Aff-Wild2 target, the rule-based method achieves the best average F1-score (25.4), driven by high and stable performance on happy (70.5 ± 1.1*). This suggests that when a source domain is sufficiently rich, the underlying AU detectors can generalize more effectively, which in turn empowers the knowledge-based reasoning layer.

The value of the ensemble Across these challenging transfer tasks, the results demonstrate that neither a purely data-driven nor a purely knowledge-driven approach is universally optimal. Each has class-specific strengths, such as the rule-based systems competence on subtle emotions like sadness, and weaknesses, like its difficulty with fear. The true value of the ensemble lies in its ability to provide a balanced and robust performance profile. By mitigating the most severe weaknesses of its individual components, an ensemble can offer a more reliable and consistent solution. This pragmatic combination is crucial for applied intelligence, where avoiding catastrophic failure is often more important than maximizing peak performance.

Beyond predictive performance, this ensemble structure is fundamental to achieving the goals of trustworthy AI. The disagreement between the data-driven model's output and the knowledge-driven rule-based prediction serves as a direct, quantifiable signal for prediction uncertainty.

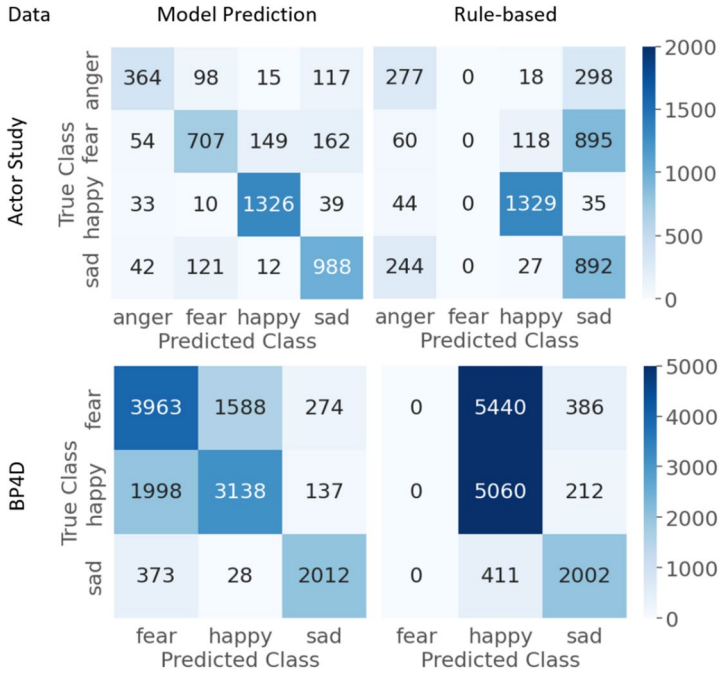

Fig. 4 Within-dataset comparison of model-based vs. fuzzy rule-based FE predictions on Actor Study (top) and BP4D (bottom). Confusion matrices show similar behavior for happy and sad, and weaker rule-based performance for fear due to a stricter three-AU rule ($1 \wedge 2 \wedge 4$)

Simultaneously, the interpretable nature of the rules provides a window into one component of the decision-making process.

5.4.2 Validating disagreement-based risk as a reliability signal

Having defined our metric for disagreement-based risk, we evaluate whether it serves as a reliable signal for prediction accuracy. Figure 5 shows calibration-style plots in a cross-dataset setting (Actor Study → Aff-Wild2): for each FE class, samples are grouped into equal-frequency bins (left to right = increasing risk/uncertainty), and the bar height is the mean accuracy within each bin. The top row uses an entropy-based uncertainty baseline (model softmax); the bottom row uses our disagreement-based risk (model vs. rule head).

Large disagreement indicates a divergence between the learned patterns and the AU→FE rules. Such conflicts can arise from noise in AU predictions, a mis-specified or incomplete rule, or inherent ambiguity in the face. However, from the perspective of a human-in-the-loop system, the source is less important than the signal itself: a conflict has been detected that needs closer inspection.

Across classes, our disagreement-based risk exhibits the desired inverse trend, where a higher risk accompanies a lower accuracy - performing comparably to the established entropy baseline. This supports the use of our proposed disagreement-based risk as a lightweight indicator of potential unreliability.

Table 10 Cross-dataset evaluation on Aff-Wild2 (val) and CK+ for different FE prediction methods

Predictor	Anger	Fear	Happy	Sad	Av.
Actor Study → Aff-Wild2					
Rule-based	**10.7** ± **5.2***	0.0 ± 0.0	51.8 ± 5.6	**25.8** ± **9.9***	22.1 ± 2.5
Model-based	8.9 ± 3.3	**4.0** ± **3.7***	54.6 ± 8.1	23.8 ± 12.0	**22.8** ± **1.1**
Ensemble	8.5 ± 3.4	2.9 ± 3.1	**55.2** ± **6.7***	23.3 ± 9.9	22.5 ± 0.8*
Actor Study → CK+					
Rule-based	19.3 ± 17.6	0.0 ± 0.0	64.1 ± 8.7	**37.4** ± **8.0***	30.2 ± 4.0
Model-based	40.1 ± 12.2	**25.2** ± **8.2***	65.4 ± 4.5	22.7 ± 10.2	38.4 ± 4.8
Ensemble	**40.9** ± **12.1***	23.8 ± 10.7	64.6 ± 4.7	26.4 ± 7.4	**38.9** ± **3.7***
[69]	20.3	16.0	**81.4**	26.0	35.9
BP4D → Aff-Wild2					
Rule-based	–	0.0 ± 0.0	**70.5** ± **1.1***	**5.8** ± **7.9***	**25.4** ± **2.3***
Model-based	–	**13.0** ± **8.2***	51.5 ± 21.8	0.3 ± 0.4	21.6 ± 7.2
Ensemble	–	8.2 ± 8.2	62.3 ± 12.4	0.1 ± 0.1	23.5 ± 1.5

Results are in F1-score. Values are mean ± SD over the 3 source-fold models evaluated on the fixed target set. **Bold** marks the highest mean in each column. An asterisk (∗) marks the smallest SD among methods whose mean lies within 0.5 points of the best mean in that column (stability among top performers). ResNet50+GRU results of [69] are from [15]

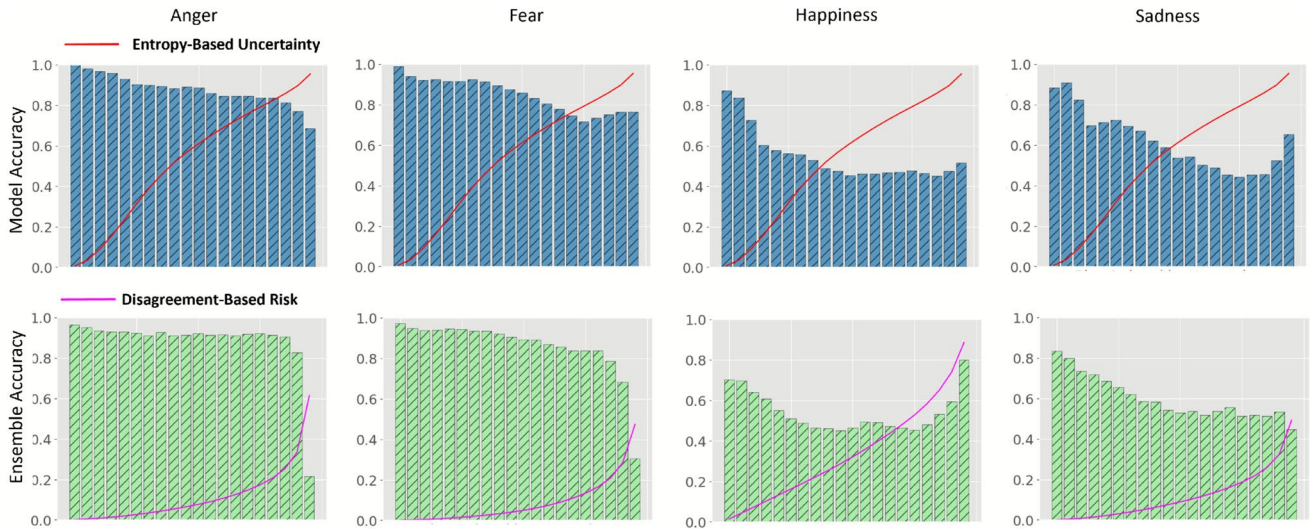

Fig. 5 Calibration plots showing the relationship between prediction accuracy and two metrics in a cross-dataset scenario (Actor Study → Aff-Wild2). Top row shows entropy. Bottom row shows our proposed metric for disagreement-based risk. Bins are ordered by increasing risk/uncertainty from left to right. Both metrics demonstrate the desired inverse relationship, where a higher value correlates with lower accuracy, validating their utility as signals of potential prediction unreliability

Importantly, the signal is also actionable. High disagreement localizes the issue to specific AU-level rules, enabling case-wise audit: which rules fired, with what support, and how they differ from the models distribution. This complements entropy (which lacks semantics), helping practitioners prioritize review and perform targeted adjustments to rules or thresholds without retraining.

5.5 Comparison with Competitive Approaches

This section compares our regularized multi-task network (FMC-Net) against relevant competitive approaches on three datasets. We compare performance across different datasets, acknowledging nuances and methodological differences that may influence comparative outcomes.

Actor study Due to the limited public availability of Actor Study, few deep baselines exist under comparable protocols. We therefore include two widely used toolkits as contextual baselines: OpenFace [70] and AUReader [14]. Both are non-deep pipelines (feature engineered) and report results on a fixed actor split (train actors 1–10, test 11–20), whereas our method is evaluated with k-fold cross-validation. Absolute AUCs are thus not strictly comparable across protocols but remain informative for context.

Table 11 presents AU detection performance (AUC). Our FMC-Net result uses the R_{task} PCC/L2 regularization variant, averaged over all folds. The results show that our deep, constraint-guided multi-task model attains a higher average AU AUC (86.1) than AUReader (73.2) and OpenFace (77.8), illustrating the benefit of our constrained learning approach. FE results of other approaches are not available for this dataset.

Table 11 Comparison with competitive approaches on Actor Study

Methods	Tasks	AU Av.	FE Av.
AUReader [14]	AU	73.2	-
OpenFace [70]	AU, other	77.8	-
FMC-Net (ours)	AU, FE	**86.1**	94.0

Our results (FMC-Net) are averaged over all folds. Results for OpenFace are from [14]. Results are in AUC. Av. is the average AUC across classes. Best results are **bold**

Aff-Wild2 Aff-Wild2 is an in-the-wild video dataset with substantial temporal and contextual variability. To avoid confounds from large external FE pretraining (e.g., AffectNet), we restrict the comparison to models trained on Aff-Wild2 or similar single/multi-dataset strategies without incorporating large external FE-specific datasets. We include single-task FE approaches (EfficientNet [39]; SSL+DTM [17]) and multi-task AU+FE approaches (a transformer-based joint model [42]; MAE-style pretraining with multi-task heads [4]). These approaches cover dominant deep approaches (frame-based vs. temporal, single-task vs. multi-task). Compared to end-to-end deep baselines that couple tasks implicitly via shared features, our FMC-Net variant explicitly injects structured AU–FE knowledge at training time through a soft constraint loss.

Table 12 reports FER results (F1) for the happy and sad classes, which have sufficient frames for k-fold training and evaluation. Our FMC-Net uses the best-performing

Table 12 Comparison with competitive approaches on Aff-Wild2 for FER

Methods	Tasks	Happy	Sad	Av.
EfficientNet [39]	FE	47.7	46.1	46.9
SSL+DTM [17]	FE	45.2	49.0	47.1
Zhang et al. [42]	AU, FE	56.6	28.7	42.6
MAE [4]	AU, FE, VA	59.6	65.6	53.7
FMC-Net(ours)	AU, FE	**64.6**	**78.6**	**71.6**

Results are in F1-score. Best results are **bold**

regularized multi-task configuration from Table 8. Our constraint-guided model outperforms both single-task and multi-task baselines on happy, sad, and the average, with especially strong gains on sad. This is consistent with the hypothesis that aligning training with AU-driven structure helps the harder, less frequent class.

BP4D For AU detection on BP4D, SEV-Net [44] is a strong single-task reference that embeds semantic descriptions with visual features. We further include informed single-task AU methods such as CorrLoss [21], which regularizes AU co-occurrence structure, and CISNet [18], which incorporates causal priors as well as multi-task approaches such as MAL [20] and JAO [28]. Approaches in multi-task learning adaptively share parameters (MAL) or select relevant auxiliary tasks to optimize related outputs such as optical flow prediction (JAO). MAL dp not publish their results on the FE task. Overall, these baselines are end-to-end deep models, where none employ the explicit AU–FE constraint loss used in FMC-Net.

Table 13 shows AU detection (macro F1, averaged over folds) using the R_{task} CCC/L2 configuration from Table 8. FMC-Net matches or exceeds the strongest baselines, achieving 66.2 macro F1 versus 66.1 for SEV-Net, and surpassing other informed single- and multi-task approaches, indicating that constraint-guided training is competitive for AU detection.

Table 14 reports FER results (macro F1) on three classes (fear, happy, sad) chosen for data availability and comparability. FMC-Net (same R_{task} CCC/L2 configuration) outperforms ResNet18+Smote [40] on the full dataset (72.7 vs. 70.0). For comparability with uGMM-MIK [71], which evaluated a subject subset, we additionally report our best

Table 13 Comparison with competitive approaches on BP4D for AUs

Methods	Tasks	AU Av.	FE Av.
SEV-Net [44]	AU	**66.1**	-
CorrLoss [21]	AU	63.8	-
CISNet [18]	AU	64.4	-
MAL [20]	AU, FE	63.3	-
JAO [28]	AU, Optical Flow	64.7	-
FMC-Net (ours)	AU, FE	**66.2**	72.7

Results are in macro F1-score. Best results are **bold**

Table 14 Comparison with competitive approaches on BP4D for FER

Methods	Fear	Happy	Sad	Av.
ResNet18+Smote [40]	-	-	-	70.0
FMC-Net (ours)	68.2	62.5	87.4	**72.7**
uGMM-MIK [71], subset	71.6	68.3	83.7	74.5
FMC-Net (ours), subset	**72.2**	**71.2**	**93.7**	**79.0**

Results are in F1-score. Best results are **bold**

randomly split fold (subject IDs F006, F007, F010, F012, F014, F022, M001, M002, M005, M007, M009, M010, M014, M018), where FMC-Net attains a higher average macro F1 (79.0 vs. 74.5). This indicates that our constraint-guided multi-task training effectively addresses FER.

6 Discussion

FMC-Net was designed to unite three key goals for applied AI: adaptability via constraint-guided training and logical rules, transparency via a rule-based fuzzy layer, and practical risk management via an ensemble with a disagreement signal. Across three heterogeneous datasets and cross-dataset tests, our results validate this multi-faceted approach.

Our results yield three takeaways for practice. (1) Constraints consistently improve performance, especially for FE: The best-performing models in nearly every metric are constrained MT variants, with the FE task benefiting most, notably on in-the-wild data.(2) A trade-off exists between peak performance and stability. PCC-based constraints often yield the highest mean scores, whereas CCC-based variants frequently demonstrate lower variance. This is a crucial consideration for deployment, as selecting solely by the highest mean can be misleading on datasets with high variance (e.g., due to a small subject number); a stability-aware perspective is beneficial. (3) Interpretable disagreement has operational value: The conflict between the data-driven and rule-based predictions offers a human-legible insight that a specific sample predictions may be high-risk and require closer inspection.

To address the cost and scalability of the rule layer, we use a compact, reusable, soft rule set derived from EMFACS [24] and apply the same default rules across datasets. There is no per-dataset manual engineering. The training-time prior is likewise soft: a fixed FE→AU expectation or co-occurrence estimate used as a correlation/consistency regularizer. Because both mechanisms are soft (regularizer; fuzzy rules), imperfect or incomplete knowledge leads to graceful degradation rather than failure. When domain adaptation is warranted (e.g., clinical deployment), edits can be local (a handful of rule weights/thresholds) and bootstrapped from AU–FE co-occurrence statistics to minimize expert effort.

We also acknowledge and mitigate threats to the validity of our results. (1) First, to address potential label noise from sequence-level annotations (Actor Study, BP4D), we apply an automated denoising filter. (2) Our results reveal sensitivity to rare/strict patterns (e.g., fear). Our ensemble helps mitigate these failures, but this highlights that rules may need to be tuned for certain target classes. (3) Regarding the cost of knowledge specification, we acknowledge that while it is a real cost, it is also a feature for safety-critical applications. (4) We frame the disagreement metric as a pragmatic risk signal (model/knowledge conflict), not a calibrated probabilistic uncertainty measure. Broader comparisons with alternative risk proxies are a useful direction for future work.

To summarize, under realistic protocols with limited folds and heterogeneous data, FMC-Net offers a balanced package for applied AI: constraints to improve performance (especially for FE), CCC variants to stabilize training, and a transparent second layer (rules plus disagreement) to expose when predictions may be unreliable. This combination is attractive for settings where both accuracy and transparency matter.

7 Conclusion

We presented FMC-Net, a human-guided FER framework that couples data-driven learning with psychology-grounded structure via two complementary constraint layers. A hierarchy-aware, inter-task correlation/consistency regularizer shapes the shared representation during multi-task training, and a rule-based fuzzy layer provides a transparent AU→FE pathway. An ensemble over the model and rule heads exposes a disagreement-based risk signal. Across controlled and in-the-wild datasets, constraint-guided training improves multi-task performance and generalization, while the rule pathway yields per-decision attributions and actionable risk cues.

FMC-Net is model-agnostic and readily extensible, opening several promising directions for future work. Methodologically, the framework could be extended with (1) masked/weighted constraints to handle untrusted data pairs; (2) soft, weighted rule relaxation to reduce brittleness; and (3) learnable rule weights for lightweight domain adaptation.

Broader future work should focus on usability and validation. This includes developing intuitive authoring tools to make knowledge specification accessible to non-experts, and validating the framework across a wider range of structured, multi-output problems, from fine-grained pain estimation and compound emotion to industrial quality inspection. A systematic comparison of our disagreement-based risk score against other reliability proxies would also be a valuable contribution.

Overall, FMC-Net offers a practical path toward AI systems that balance accuracy with transparency and controllability, aligning with principles of trustworthy AI and regulatory expectations (e.g., the EU AI Act) for high-risk applications such as biometric categorization and emotion recognition.

Acknowledgements The authors thank Betelhem Nebebe for their contribution to the development of the multi-task models. They thank Robert Obermeier, Jan Adelhardt, and Anna Moeller for their efforts in maintaining the infrastructure and pre-processing the data.

Author Contributions Ines Rieger: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing–original draft, Writing–review & editing, Visualization, Project administration. Jaspar Pahl: Conceptualization, Methodology, Software, Data curation, Writing–review & editing. Ute Schmid: Conceptualization, Methodology, Supervision, Writing–review & editing.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The datasets analysed during the current study are publicly available datasets: Actor Study [63], BP4D-Spontaneous Dataset [64], Aff-Wild2 Database [65], and the extended Cohn-Kanade (CK+) database [66]. Information on how to access these datasets is available through the corresponding citations.

Declarations

Competing interests The authors declare they have no competing interests. This article has never been submitted to more than one journal for simultaneous consideration. This article is original.

Consent for publication All authors have read and approved the final manuscript for publication.

Ethical Standards This research relies solely on the analysis of publicly available datasets, as detailed in the Data Availability Statement. No new studies involving human participants or animals were conducted by the authors. The original collection and distribution of these datasets have been reported by their creators to comply with relevant ethical regulations and informed consent procedures.

References

1. Picard RW (2000) Affective Computing. MIT press, Cambridge, MA
2. Li S, Deng W (2020) Deep facial expression recognition: A survey. IEEE Trans Affect Comput 13(3):1195–1215
3. Kollias D, Tzirakis P, Baird A, Cowen A, Zafeiriou S (2023) Abaw: Valence-arousal expression action unit & emotional intensity estimation challenges. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5888–5897
4. Zhang W, Ma B, Qiu F, Ding Y (2023) Multi-modal facial affective analysis based on masked autoencoder. In: Proceedings of IEEE/CVF conference on computer vision and pattern recognition, pp 5792–5801
5. Dehshibi MM, Masip D (2024) Bee-net: A deep neural network to identify in-the-wild bodily expression of emotions. arXiv preprint arXiv:2402.13955
6. Kim J, Lee D (2023) Facial expression recognition robust to occlusion and to intra-similarity problem using relevant subsampling. Sensors 23(5):2619
7. Ekman P, Friesen WV (1978) Facial action coding systems. Consulting Psychologists Press Palo Alto CA
8. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput Surv (CSUR) 51(5):1–42
9. Schmid U (2023) Trustworthy artificial intelligence: Comprehensible transparent and correctable. In: Introduction to digital humanism: A Textbook. Springer
10. Commission E.: Artificial Intelligence Act (2024). https://artificialintelligenceact.eu
11. Alipour M, Moghaddam MT, Vaidhyanathan K, Kjærgaard MB (2023) Emoticontrol: Emotions-based control of user-interfaces adaptations. Proc ACM Human-Comput Interact 7(EICS):1–29
12. Gamage G, De Silva D, Mills N, Alahakoon D, Manic M (2024) Emotion aware: An artificial intelligence framework for adaptable robust explainable and multi-granular emotion analysis. J Big Data 11(1):93
13. Du Z, Wu S, Huang D, Li W, Wang Y (2021) Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. IEEE Trans Affect Comput 12(03):565–578
14. Seuss D, Hassan T, Dieckmann A, Unfried M, Scherer K, Mortillaro M, Garbas J-U (2023) Automatic estimation of action unit intensities and inference of emotional appraisals. IEEE Trans Affect Comput 14(2):1188–1200
15. Pahl J, Rieger I, Möller A, Wittenberg T, Schmid U (2022) Female, white 27? bias evaluation on data and algorithms for affect recognition in faces. In: Proceedings of the ACM conference on fairness accountability and transparency, pp 973–987
16. Ren Z, Ortega J, Wang Y, Chen Z, Guo Y, Yu SX, Whitney D (2024) Veatic: Video-based emotion and affect tracking in context dataset. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 4467–4477
17. Yu J, Cai Z, Li R, Zhao G, Xie G, Zhu J, Zhu W, Ling Q, Wang, L, Wang C 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
18. Chen Y, Chen D, Wang T, Wang Y, Liang Y (2022) Causal intervention for subject-deconfounded facial action unit recognition. In: Proceedings of AAAI conference on artificial intelligence, vol 36, pp 374–382
19. Rieger I (2024) Investigating the regularization of deep neural networks for affect recognition with relevance-guided local explanations. In: International conference on artificial intelligence: Methodology systems and applications, Springer, pp 122–127
20. Li Y, Shan S (2023) Meta auxiliary learning for facial action unit detection. IEEE Trans Affect Comput 14:2526–2538
21. Rieger I, Pahl J, Finzel B, Schmid U (2022) Corrloss: Integrating co-occurrence domain knowledge for affect recognition. In: Proceedings of the international conference on pattern recognition, IEEE, pp 798–804
22. Dash T, Chitlangia S, Ahuja A, Srinivasan A (2022) A review of some techniques for inclusion of domain-knowledge into deep neural networks. Sci Rep 12(1):1040
23. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Pers Soc Psychol 17(2):124
24. Friesen WV, Ekman P et al (1983) Emfacs-7: Emotional facial action coding system
25. Ekman P, Friesen WV, Hager JC (2002) Facial action coding system: The manual on cd rom. A Human Face Salt Lake City 77–254
26. Kunz M, Lautenbacher S (2014) The faces of pain: A cluster analysis of individual differences in facial activity patterns of pain. Eur J Pain 18(6):813–823
27. Hassan T, Seuß D, Wollenberg J, Weitz K, Kunz M, Lautenbacher S, Garbas J-U, Schmid U (2019) Automatic detection of pain from facial expressions: A survey. IEEE Trans Pattern Anal Mach Intell 43(6):1815–1831
28. Shao Z, Zhou Y, Li F, Zhu H, Liu B (2024) Joint facial action unit recognition and self-supervised optical flow estimation. Pattern Recogn Lett 181:70–76
29. Wang S, Chang Y, Wang C (2023) Dual learning for joint facial landmark detection and action unit recognition. IEEE Trans Affect Comput 14(02):1404–1416
30. Li X, Deng W, Li S, Li Y (2023) Compound expression recognition in-the-wild with au-assisted meta multi-task learning. In: Proceedings of IEEE/CVF conference on computer vision and pattern recognition, pp 5734–5743
31. Siebers M, Schmid U, Seuß D, Kunz M, Lautenbacher S (2016) Characterizing facial expressions by grammars of action unit sequences-a first investigation using abl. Inf Sci 329:866–875
32. Posner J, Russell JA, Peterson BS (2005) The circumplex model of affect: An integrative approach to affective neuroscience cognitive development and psychopathology. Dev Psychopathol 17(3):715–734
33. Calvo RA, D'Mello S (2010) Affect detection: An interdisciplinary review of models methods and their applications. IEEE Trans Affect Comput 1(1):18–37
34. Schuller B, Batliner A, Bergmann K, Steidl S, Vogt T (2018) Speech emotion recognition: Two decades in a nutshell benchmarks and ongoing trends. Speech Commun 87:180–196
35. Medjden S, Ahmed N, Lataifeh M (2020) Adaptive user interface design and analysis using emotion recognition through facial expressions and body posture from an rgb-d sensor. PLoS ONE 15(7):0235908
36. Santhanaraj KK, MM R (2021) A survey of assistive robots and systems for elderly care. J Enabling Technol 15(1):66–72
37. Shneiderman B (2022) Human-centered AI. Oxford University Press, Oxford UK
38. Holzinger A (2016) Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Inf 3(2):119–131
39. Savchenko A (2022) Video-based frame-level facial analysis of affective behavior on mobile devices using efficientnets. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2359–2366
40. Nourivandi T, Hinduja S, Srivastava S, Cohn JF, Canavan S (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
41. Wang C, Zeng J, Shan S, Chen X (2019) Multi-task learning of emotion recognition and facial action unit detection with adaptively weights sharing network. In: Proceedings of the IEEE international conference on image processing, IEEE, pp 56–60
42. Zhang W, Qiu F, Wang S, Zeng H, Zhang Z, An R, Ma B, Ding Y (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
43. Li G, Zhu X, Zeng Y, Wang Q, Lin L (2019) Semantic relationships guided representation learning for facial action unit recognition. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8594–8601
44. Yang H, Yin L, Zhou Y, Gu J (2021) Exploiting semantic embedding and visual feature for facial action unit detection. In: Proceedings of IEEE/CVF conference on computer vision and pattern recognition, pp 10482–10491
45. Cui Z, Song T, Wang Y, Ji Q (2020) Knowledge augmented deep neural networks for joint facial expression and action unit recognition. Proc Adv Neural Inf Process Syst 33:14338–14349
46. Valstar MF, Pantic M (2006) Biologically vs. logic inspired encoding of facial actions and emotions in video. In: 2006 IEEE international conference on multimedia and expo, IEEE, pp 325–328
47. Velusamy S, Kannan H, Anand B, Sharma A, Navathe B (2011) A method to infer emotions from facial action units. In: Proceedings of the IEEE international conference on acoustics speech and signal processing, pp 2028–2031
48. Hájek P, Godo L, Esteva F (1996) A complete many-valued logic with product-conjunction. Arch Math Logic 35:191–208
49. Badreddine S, Garcez AD, Serafini L, Spranger M (2022) Logic tensor networks. Artif Intell 303:103649
50. Schwalbe G, Wirth C, Schmid U (2022) Enabling verification of deep neural networks in perception tasks using fuzzy logic and concept embeddings. arXiv preprint arXiv:2201.00572
51. Liang C, Wang W, Miao J, Yang Y (2023) Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In: Proceedings of IEEE/CVF international conference on computer vision, pp 16197–16208
52. Morales-Vargas E, Reyes-García C, Peregrina-Barreto H (2019) On the use of action units and fuzzy explanatory models for facial expression recognition. PloS One 14
53. Vashishtha S, Gupta V, Mittal M (2023) Sentiment analysis using fuzzy logic: A comprehensive literature review. Wiley Interdiscip Rev Data Min Knowl Discov 13(5):1509
54. Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. Proc Adv Neural Inf Process Syst 30
55. Gawlikowski J, Tassi C.R.N, Ali M, Lee J, Humt M, Feng J, Kruspe, A, Triebel R, Jung P, Roscher R et al (2023) A survey of uncertainty in deep neural networks. Artif Intell Rev 56(Suppl 1):1513–1589
56. Deuschel J, Foltyn A, Roscher K, Scheele S (2024) The role of uncertainty quantification for trustworthy ai 95–115
57. Zhang W, Qiu F, Liu C, Li L, Du H, Guo T, Yu X (2024) 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
58. Geifman Y, El-Yaniv R (2017) Selective classification for deep neural networks. Adv Neural Inf Process Syst 30
59. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 770–778
60. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: Proceedings of IEEE/CVF conference on computer vision and pattern recognition, pp 248–255
61. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
62. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, pp 249–256
63. Seuss D, Dieckmann A, Hassan T, Garbas JU, Ellgring JH, Mortillaro M, Scherer K (2019) Emotion expression from different angles: A video database for facial expressions of actors shot by a camera array. In: Proceedings of the international conference on affective computing and intelligent interaction, pp 35–41. https://doi.org/10.1109/ACII.2019.8925458
64. Zhang X, Yin L, Cohn JF, Canavan S, Reale M, Horowitz A, Liu P, Girard JM (2014) Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database. Image Vis Comput 32(10):692–706
65. Kollias D, Zafeiriou S (2018) Aff-wild2: Extending the aff-wild database for affect recognition. preprint arXiv:1811.07770
66. Lucey P, Cohn J.F, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Computer society conference on computer vision and pattern recognition, IEEE, pp 94–101
67. Bradski G (2000) The opencv library. Dr. Dobb's journal: Software tools for the professional programmer 25(11):120–123
68. Reddi SJ, Kale S, Kumar S (2018)On the convergence of adam and beyond. In: International conference on learning representation
69. Deng D, Chen Z, Shi BE (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
70. Baltrusaitis T, Zadeh A, Lim YC, Morency LP (2018) Openface 2.0: Facial behavior analysis toolkit. In: Proceedings of the IEEE international conference on automatic face and gesture recognition, IEEE, pp 59–66
71. Perveen N, Roy D, Chalavadi KM (2020) Facial expression recognition in videos using dynamic kernels. IEEE Trans Image Process 29:8316–8325