Leveraging Cross-Modal Information to Reduce Gender Bias in Facial Analysis Systems

Iris Dominguez-Catena, Daniel Paternain, Aranzazu Jurio, Mikel Galar

Institute of Smart Cities (ISC)
Universidad Pública de Navarra (UPNA)

May 13, 2026

upna Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

arin Artificial Intelligence & Machine Learning Research

[0] iris.dominguez@unavarra.es

Motivation: Bias in Facial Analysis

- Facial analysis systems are ubiquitous: HCI, security, healthcare
- Well-documented demographic biases[1]:
  - Gender classification: up to 34.7% error for darker-skinned females vs. near-zero for lighter-skinned males
- **Underexplored**: intersection of demographic attributes and facial expressions
  - Gender prediction fails more on women displaying anger/fear
  - Emotion recognition shows disparate performance across genders

Goal

Mitigate cross-modal gender-emotion biases **without** retraining base models

[1] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proc. 1st Conf. Fairness Account. Transpar.* Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, Feb. 2018, pp. 77–91.

Existing Bias Mitigation Approaches

Pre-processing[1]

- Data augmentation / resampling
- Requires access to training data

In-processing[2]

- Fairness constraints in loss
- Architecture-specific

Post-processing[3]

- Calibrate predictions
- Rigid fairness definitions

Common Limitations

- Require retraining or training data access
- High computational cost
- Often trade accuracy for fairness[4]
- Model-specific solutions

[1] Angelina Wang et al. "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets". In: *Int J Comput Vis* 130.7 (July 2022), pp. 1790–1810. ISSN: 1573-1405. DOI: 10.1007/s11263-022-01625-5.

[2] Muhammad Bilal Zafar et al. "Fairness Constraints: Mechanisms for Fair Classification". In: *Proc. 20th Int. Conf. Artif. Intell. Stat.* PMLR, Apr. 2017, pp. 962–970.

[3] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* NIPS'16. Red Hook, NY, USA: Curran Associates Inc., Dec. 2016, pp. 3323–3331. ISBN: 978-1-5108-3881-9.

[4] Aditya Krishna Menon and Robert C. Williamson. "The Cost of Fairness in Binary Classification". In: *Proc. 1st*

Proposal: Exploit Complementary Biases

FairFace[5]
(Gender prediction)

- Struggles with female faces displaying anger or fear
- Emotion-dependent failures

EmoNet[6]
(Emotion recognition)

- Gender-dependent accuracy patterns
- Poor on neutral/sad for both genders

Opportunity

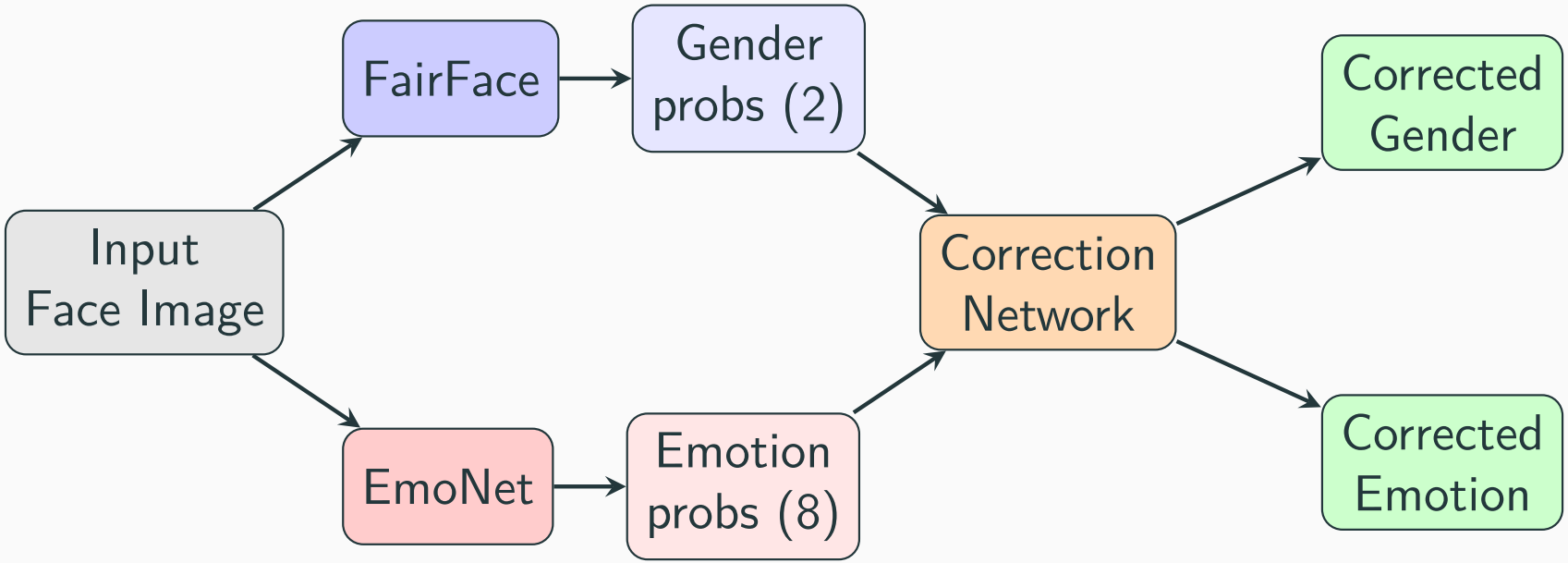
Failure modes are **complementary**: if each model had awareness of the other's predictions, they could *naturally* correct for each other's biases.

---

[6] Kimmo Karkkainen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation". In: *2021 IEEE Winter Conf. Appl. Comput. Vis. WACV*. Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 1547–1557. ISBN: 978-1-6654-0477-8. DOI: 10.1109/WACV48630.2021.00159.

[7] Antoine Toisoul et al. "Estimation of Continuous Valence and Arousal Levels from Faces in Naturalistic Conditions". In: *Nat Mach Intell* 3.1 (Jan. 2021), pp. 42–50. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00280-0.

Methodology: Cross-Modal Correction



- Concatenate probability outputs: 10-dim input (2 + 8)
- Joint optimization: correction network loss is average of gender and emotion cross-entropy losses
- Only the correction network gets trained
- We do not impose **explicit fairness constraints**: the correction emerges naturally

Correction Network Architecture

Architecture:

- Input: 10 dimensions (2 gender + 8 emotion probs)
- 3 fully connected layers, 24 neurons each
- ReLU activations, dropout ($p = 0.2$)
- Output: 10 dimensions (2 gender + 8 emotion probs), independent softmax in each set

Training:

- Adam optimizer, lr = 0.001
- 500 epochs, batch size 32
- Training time: ∼42 seconds (convergence at ∼10s)

Key Properties

- **Model-agnostic**: works with any base models
- **No retraining**: base models remain frozen
- **Lightweight**: minimal overhead
- **No fairness constraints**: bias correction emerges implicitly

Experimental Setup: Datasets

Training: **FACES**[8]

- 2,052 images, 171 subjects
- Balanced gender and age
- 6 emotions: neutral, sad, disgust, fear, anger, happy

Test: **ADFES**[9]

- 216 images, 22 subjects
- 10 female, 12 male
- Same 6 emotions

Evaluation Metrics:

- **Macro-average accuracy**: $\text{Acc}_{\text{macro}} = \frac{1}{12} \sum_{g,e} A_{g,e}$ (fairness-aware, considers gender and emotion)
- **Gender gap per emotion**: $\text{Gap}_e = |A_{\text{male},e} - A_{\text{female},e}|$
- **Average gap**: $\text{Gap}_{\text{avg}} = \frac{1}{6} \sum_e \text{Gap}_e$ (lower = fairer)

---

[9] Natalie C. Ebner, Michaela Riediger, and Ulman Lindenberger. "FACES—A Database of Facial Expressions in Young, Middle-Aged, and Older Women and Men: Development and Validation". In: *Behavior Research Methods* 42.1 (Feb. 2010), pp. 351–362. ISSN: 1554-3528. DOI: 10.3758/BRM.42.1.351.

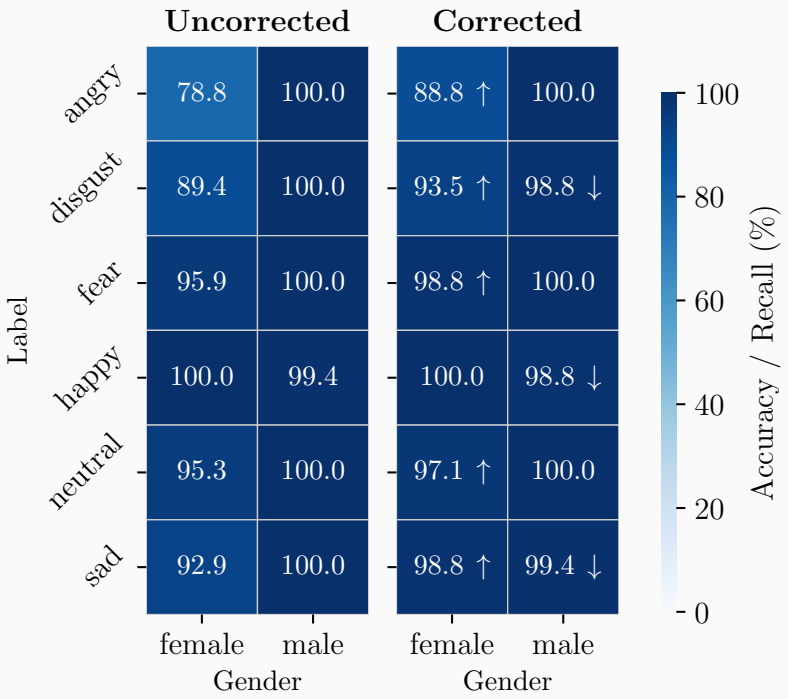
[9] Job van der Schalk et al. "Moving Faces, Looking Places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES)". In: *Emotion* 11.4 (2011), pp. 907–920. ISSN: 1931-1516. DOI: 10.1037/a0023853.

Results: Overall Performance

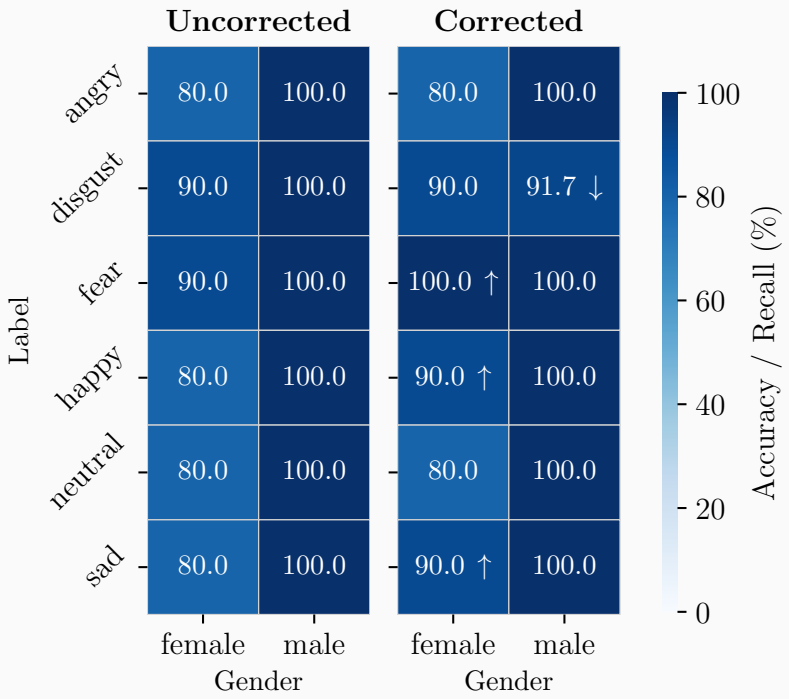
Dataset	Task	Correction	Macro-Avg Acc. (%)	Largest Gap (%)	Average Gap (%)
FACES	Gender	Uncorrected	95.98	21.18	8.04
		Corrected	**97.85**	**11.18**	**3.73**
	Emotion	Uncorrected	74.57	13.50	4.20
		Corrected	**93.91**	**5.22**	**2.62**
ADFES	Gender	Uncorrected	91.67	20.00	16.67
		Corrected	**93.47**	20.00	**10.28**
	Emotion	Uncorrected	80.69	**25.00**	13.06
		Corrected	**92.92**	40.00	**7.50**

- Emotion accuracy on unseen data: 80.7% $\rightarrow$ 92.9% (+12.2 pp)
- Average gender gap (emotion): 13.1% $\rightarrow$ 7.5% (-5.6 pp)
- Improvements transfer across datasets despite distribution shift

Results: Gender Prediction



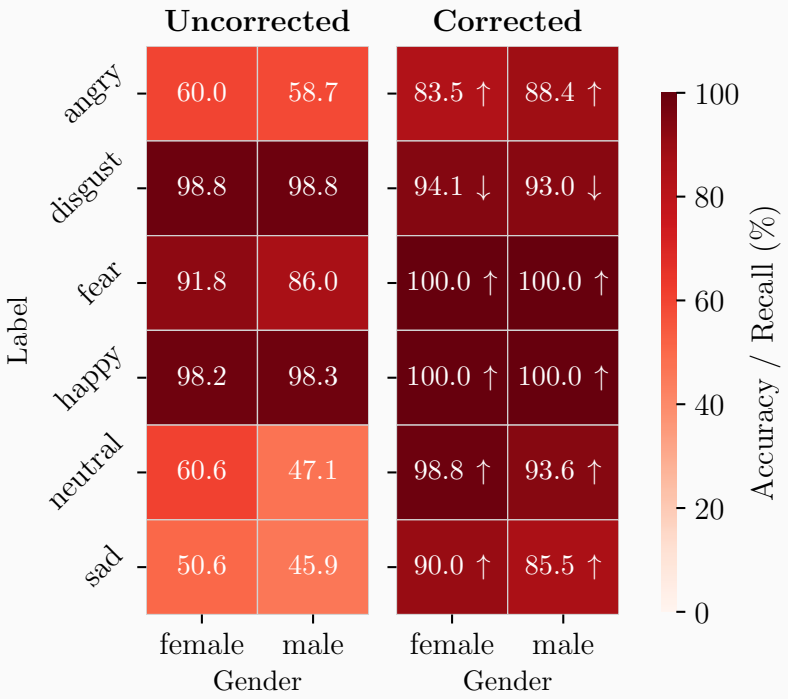
(a) FACES (training)



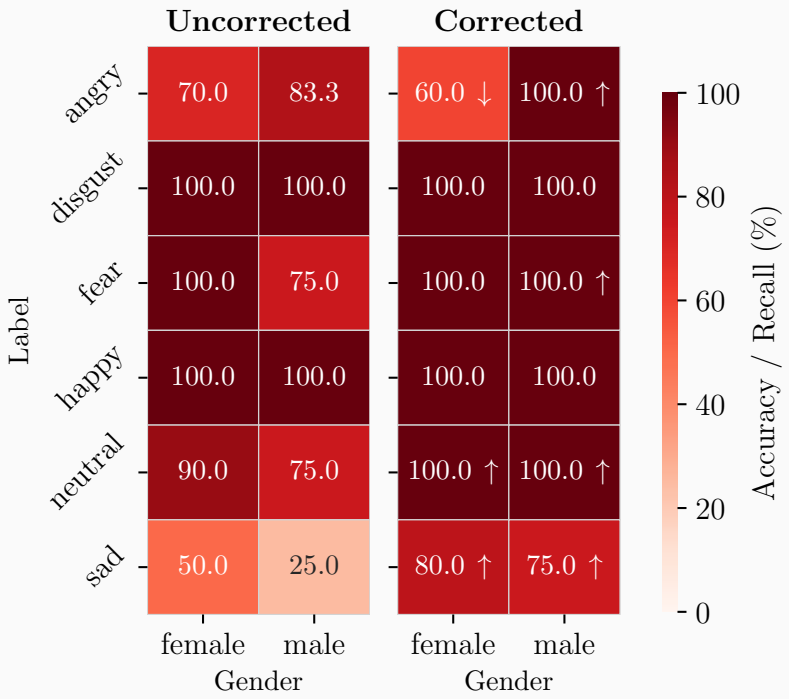
(b) ADFES (test)

- Systematic bias against female faces
- Correction reduces gaps: angry 21.2% $\rightarrow$ 11.2%, disgust 10.6% $\rightarrow$ 5.3%
- Generalizes to ADFES: sad 80% $\rightarrow$ 90%, fear 90% $\rightarrow$ 100% for females

Results: Emotion Prediction



(c) FACES (training)



(d) ADFES (test)

- Most dramatic improvements in FACES: angry, neutral, sad
- ADFES: sad improves from 50%/25% to 80%/75% (female/male)
- Trade-off: angry female drops 70% → 60% on ADFES (distribution shift)

Discussion: Why Does It Work?

Fairness Through Awareness

Cross-modal information provides **contextual signals** that disambiguate failure cases:

- Gender probs help calibrate emotion predictions for demographic subgroups
- Emotion probs help identify when gender prediction is unreliable

Key observations:

- Bias often stems from **missing contextual information**, not fundamental model limitations
- Joint awareness enables implicit bias correction without explicit fairness constraints
- Results generalize across datasets despite distribution shift
- Lightweight approach: 42s training vs. hours for full model retraining

Conclusions and Future Work

Contributions:

1. Identified cross-modal gender-emotion biases in FairFace and EmoNet
2. Proposed lightweight, model-agnostic correction framework
3. Validated on FACES + ADFES: improvements in both accuracy and fairness

Limitations & Future Directions:

- Distribution shift challenges (angry female on ADFES)
- Extend to non-binary gender, more emotions
- Apply to other attributes (age, ethnicity) and domains

Thank you! Questions?

iris.dominguez@unavarra.es