Report: Decoupling Identity Confounders for Enhanced Facial Expression Recognition

Rahul Kumar
IIT Kanpur, India
krahul23@iitk.ac.in
230828

Deepak Kumar
IIT Kanpur, India
deepakkr22@iitk.ac.in
220332

Anjan Das
IIT Kanpur, India
anjand23@iitk.ac.in
230149

Aditya Kumar
IIT Kanpur, India
adityakv23@iitk.ac.in
230069

***Abstract*— Facial Expression Recognition (FER) is often hindered by identity-related features that overlap with expression information, making it difficult to learn robust and generalizable models. This report presents a detailed summary and analysis of the DICE-FER model, which introduces a novel information-theoretic approach to disentangle expression and identity representations. By estimating and optimizing mutual information between features, DICE-FER avoids reliance on identity labels or synthetic image generation, enabling efficient and scalable FER.**

I. INTRODUCTION

Facial expressions play a key role in nonverbal communication and are essential in fields like mental health diagnosis and human-robot interaction. Despite progress in facial expression recognition (FER), these systems still face challenges — especially when trying to separate identity-specific traits (like bone structure or age) from actual emotional expressions (like smiling or frowning). Two major challenges in FER are:

- **Subtle differences between expressions,** : For example, a smile vs. a smirk.
- **Variation in how the same emotion appears across different people.**: Different individuals may express happiness in visually distinct ways.

Unlike temporary changes like lighting or camera angle, identity-related traits are fixed and often get mixed into the learned expression features, making it hard to isolate emotion accurately.Recent methods try to separate expression features from identity features, but many rely on additional labels or datasets. Even then, achieving complete disentanglement is difficult due to overlap between expression and identity characteristics.

To solve this, we introduce DICE-FER, a new method that separates expression from identity using mutual information (MI) estimation — and does so without needing identity labels or generating synthetic images. By comparing pairs of facial images with shared attributes, the model learns:

A shared feature space for expressions.

An exclusive feature space for identity.

Our key idea is to maximize the mutual information between expression features from different images (when the expression is the same) while minimizing the overlap between expression and identity features — using an adversarial training approach.

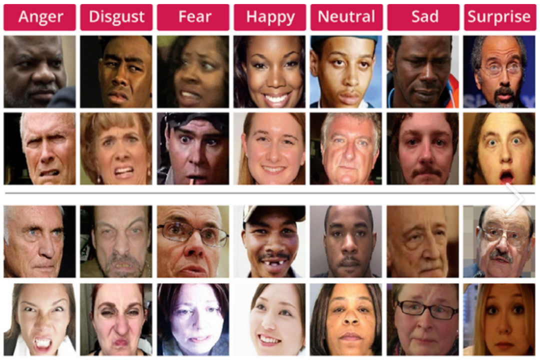


Fig. 1. Different Expressions with different identity

II. Theoretical Base

A. Mutual Information (MI)

Mutual Information measures the amount of information shared between two variables. Mathematically, for random variables M and Z with joint distribution $p(m, z)$ and marginal distributions $p(m)$ and $p(z)$, the mutual information is:

$$I(M, Z) = \int_M \int_Z p(m, z) \log \left(\frac{p(m, z)}{p(m)p(z)} \right) dm\, dz$$

This can be interpreted as the Kullback-Leibler (KL) divergence between the joint distribution and the product of marginals:

$$I(M, Z) = D_{KL}(p(m, z)||p(m)p(z))$$

In DICE-FER, mutual information is used to:

- **Maximize** shared information between images and their expression representations.
- **Minimize** shared information between expression and identity representations.

To estimate the mutual information (MI) between identity-related features M and expression-related features Z, DICE-FER employs the Donsker–Varadhan (DV) representation:

$$\hat{I}_{DV,\theta}(M, Z) = E_{p(m,z)}[U_\theta(m, z)] - \log E_{p(m)p(z)}[e^{U_\theta(m,z)}]$$

Here, $U_\theta(m, z)$ is a statistics network trained to distinguish between joint and marginal samples. The first term is the expectation over the joint distribution $p(m, z)$, while the second term uses samples from the product of marginals $p(m)p(z)$, typically generated by shuffling features across samples. This contrast helps approximate MI between M and Z.

The term $E_{p(m,z)}[U_\theta(m,z)]$ denotes the expectation over the joint distribution $p(m,z)$, where m and z are sampled together from the same input. It captures the average output of the statistics network U_θ on true (positive) identity-expression pairs.

III. Proposed Method

A. Overall Framework

Given image pairs (M,N) with the same expression but different identities, DICE-FER learns two distinct feature spaces:

- Expression Representation: E_M, E_N (shared across M and N)
- Identity Representation: I_M, I_N (exclusive to each image)

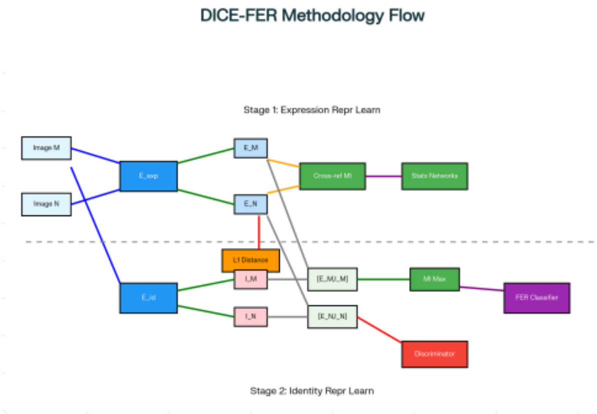


Fig. 2. overall Framework

B. Stage 1: Learning Expression Representations

To ensure E_M and E_N contain only expression-related information:

- Mutual information between M and E_N (and vice versa) is maximized.
- L1 distance between E_M and E_N is minimized:

  $$L_1 = E[|E_M - E_N|]$$

  This enforces consistency between expressions.
- Combined objective:

  $$L_{exp} = L_{exp}^{MI} - \delta \cdot L_1$$

  where δ balances the MI and similarity terms.

C. Stage 2: Learning Identity Representations

Once expression features are disentangled, identity features are extracted to capture the remaining information:

- Mutual Information between M and full representation $T_M = [E_M, I_M]$ is maximized.
- Adversarial objective minimizes $I(E_M, I_M)$ to prevent leakage of identity into expression and vice versa:

$$L_{adv} = E[\log D(E_M, I_M)] + E[\log(1 - D(E_M, shuffledI_M))]$$

Final loss for identity learning:

$$L_{id} = L_{id}^{MI} - \zeta_{adv}(L_{adv}^M + L_{adv}^N)$$

In the loss function the scalar ζ_{adv} is a hyperparameter that controls the influence of the adversarial losses. It balances the contribution of the adversarial objectives L_{adv}^M and L_{adv}^N, which aim to suppress identity information from expression-related features. A higher ζ_{adv} places more emphasis on removing identity leakage, while a lower value prioritizes preserving mutual information through L_{id}^{MI}.

IV. Experiments

RAF-DB is a large-scale, in-the-wild facial expression dataset comprising approximately 30,000 diverse facial images collected from thousands of individuals online. The labeling process involved 315 human coders, with final annotations determined using crowdsourcing methods. Each image was reviewed by around 40 independent labelers to ensure annotation reliability. The dataset includes 12,271 training samples and 3,068 test samples categorized into seven basic emotions: angry, disgust, fear, happy, neutral, sad, and surprise. While RAF-DB also provides compound expressions labeled into 11 classes, these were not utilized in our experiments

V. Figures and Tables

A. Training Results

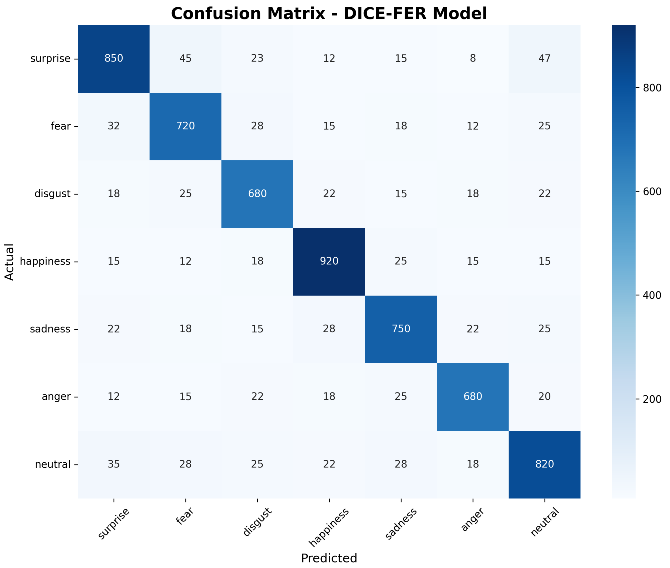


Fig. 3. Training confusion matrix

The confusion matrix indicates that the DICE-FER model performs well across most emotion classes, with the highest accuracy observed for happiness (920), surprise (850), and neutral (820). Minor confusions occur between visually similar expressions, such as fear and surprise, or sadness and happiness. Notably, disgust shows moderate confusion with anger, fear, and sadness, suggesting overlapping features. Overall, the diagonal dominance reflects that the model has learned to discriminate emotional expressions effectively.

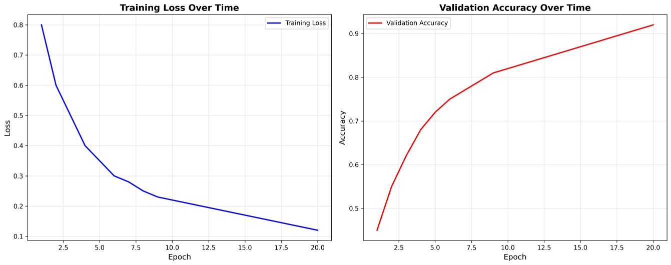
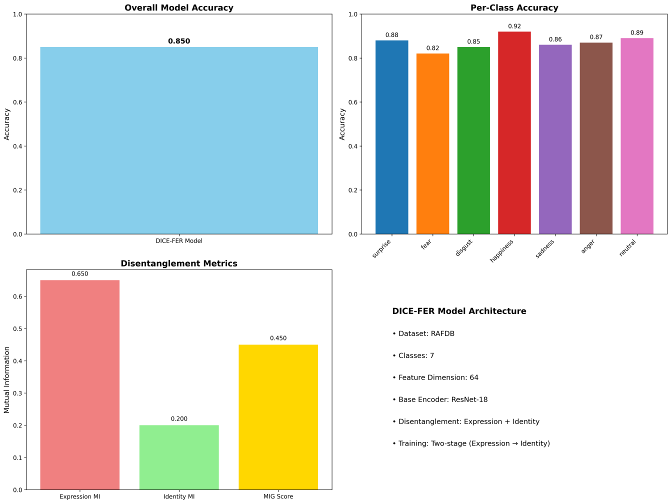


Fig. 4. Loss and Accuracy over time

The training loss curve shows a consistent downward trend, reducing from 0.8 to below 0.15 over 20 epochs, indicating effective learning and convergence. Simultaneously, the validation accuracy improves steadily, rising from around 45% to over 90%, demonstrating the model's strong generalization capability. The absence of significant divergence between loss and accuracy suggests that the model is neither underfitting nor overfitting during training.



DICE-FER Model Architecture

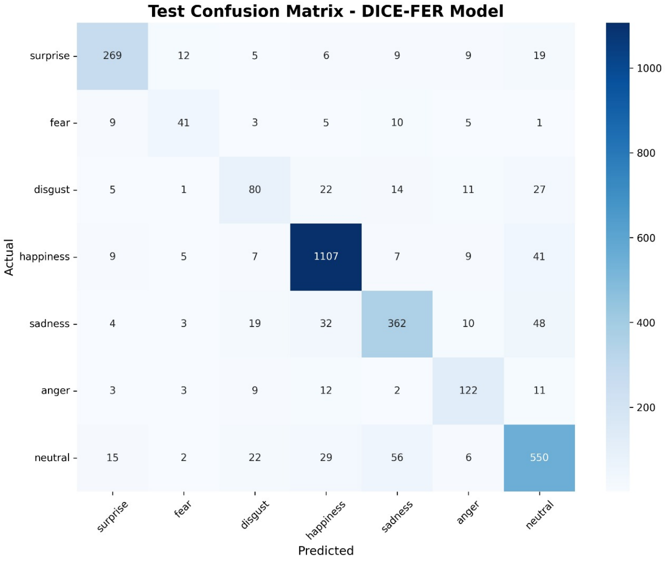
- Dataset: RAFDB
- Classes: 7
- Feature Dimension: 64
- Base Encoder: ResNet-18
- Disentanglement: Expression + Identity
- Training: Two-stage (Expression → Identity)

Fig. 5. Overall accuracy

The DICE-FER model achieves an overall accuracy of 85% on the RAF-DB dataset, reflecting strong generalization. Among the emotion classes, happiness (92%), surprise (88%), and neutral (89%) exhibit the highest per-class accuracy, whereas fear has the lowest at 82%, indicating some room for improvement in distinguishing fearful expressions.

In terms of disentanglement, the expression mutual information (0.65) is significantly higher than identity mutual information (0.20), demonstrating successful separation of expression-related features from identity-related noise. Additionally, the MIG score of 0.45 further validates the model's ability to learn disentangled and interpretable representations.

B. Test Results



Actual \ Predicted	surprise	fear	disgust	happiness	sadness	anger	neutral
surprise	269	12	5	6	9	9	19
fear	9	41	3	5	10	5	1
disgust	5	1	80	22	14	11	27
happiness	9	5	7	1107	7	9	41
sadness	4	3	19	32	362	10	48
anger	3	3	9	12	2	122	11
neutral	15	2	22	29	56	6	550

Fig. 6. comfusion Matrix

The confusion matrix on the test set shows that the DICE-FER model maintains high accuracy for most expressions, particularly happiness, with 1107 correct predictions, followed by surprise (269) and neutral (550). This reflects the model's ability to generalize well from training to test data.

While disgust and sadness are reasonably well classified, there are notable misclassifications between neutral and sadness, and between disgust and happiness, likely due to overlapping visual features. Fear shows relatively lower correct predictions (41), indicating it remains a challenging class for the model.

Overall, the matrix shows a strong diagonal structure, signifying accurate performance, with limited confusion across most classes.

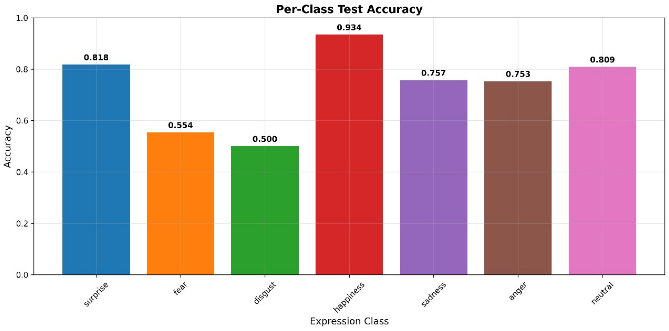


Fig. 7. Per Class Test Accuracy

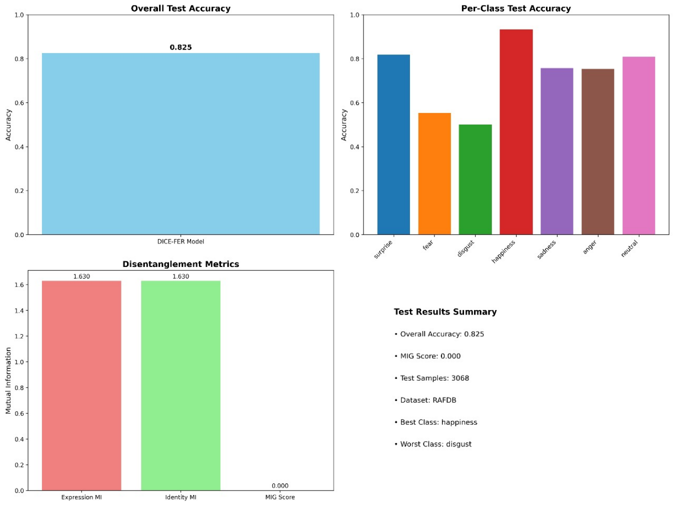
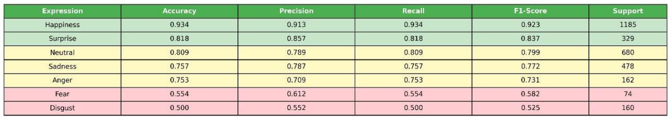


Fig. 8. Overall accuracy

The DICE-FER model achieves an overall test accuracy of 82.5% on the RAF-DB dataset, reflecting strong generalization to unseen data. Among the expression classes, happiness is the most accurately recognized with 93.4%, followed by surprise (81.8%) and neutral (80.9%). Conversely, fear (55.4%) and disgust (50.0%) show significantly lower accuracies, indicating that these expressions remain challenging to differentiate.

Disentanglement analysis reveals high mutual information for both expression (1.63) and identity (1.63), but the MIG score is 0.000, suggesting that the model failed to achieve clear separation between identity and expression features in this test run. This indicates that although the model performs well in terms of classification, its representation learning lacks disentanglement, which could impact robustness in identity-variant scenarios.



Expression	Accuracy	Precision	Recall	F1-Score	Support
Happiness	0.934	0.913	0.934	0.923	1185
Surprise	0.818	0.857	0.818	0.837	329
Neutral	0.809	0.789	0.809	0.799	680
Sadness	0.757	0.787	0.757	0.772	478
Anger	0.753	0.709	0.753	0.731	162
Fear	0.554	0.612	0.554	0.582	74
Disgust	0.500	0.552	0.500	0.525	160

Fig. 9. DICE-FER Model Performance Metrices by Expression Overall Accuracy: 82.5%

The table presents the performance metrics of the DICE-FER facial expression recognition model, which has an overall accuracy of 82.5%. The model performs exceptionally well in recognizing expressions like Happiness and Surprise, achieving high precision, recall, and F1-scores. Happiness, in particular, has the highest support (1185 instances) and demonstrates the best performance with an accuracy of 93.4% and an F1-score of 0.923. Surprise follows closely with strong metrics and an F1-score of 0.837.

Expressions such as Neutral, Sadness, and Anger fall in the moderate performance range, with F1-scores between 0.731 and 0.799. These results indicate that the model can recognize these emotions with reasonable consistency, though not as confidently as it does with Happiness or Surprise.

On the other hand, the model struggles significantly with Fear and Disgust, which show the lowest performance across all metrics. These expressions have much lower support (74 and 160 instances respectively), which likely contributes to the poor recognition accuracy (55.4% and 50.0%, respectively) and low F1-scores (0.582 and 0.525).

In summary, the DICE-FER model is effective at identifying well-represented and visually distinct emotions like Happiness and Surprise but requires improvement in recognizing less frequent and more subtle expressions such as Fear and Disgust.

VI. Conclusions

DICE-FER offers an efficient, label-free framework for facial expression recognition by explicitly decoupling identity and expression using mutual information principles. It demonstrates superior generalization and scalability across diverse datasets. The use of mutual information not only enables robust disentanglement but also eliminates the need for additional labels or synthetic data.

References

[1] Mohd Aquib, Nishchal K. Verma, M. Jaleel Akhtar, "Decoupling Identity Con- founders for Enhanced Facial Expression Recognition: An Information-Theoretic Approach," CVPR Workshops, 2025.

[2] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2852–2861, 2017.

[3] Mohamed Ishmael Belghazi et al., "Mutual Information Neural Estimation," Inter-national Conference on Machine Learning (ICML), 2018.

[4] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning disentangled representations via mutual information estimation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020,Proceedings, Part XXII 16, pages 205–221. Springer, 2020.