Towards Scalable and Ecological Methods for Early Social Development Research: Validating Automated Facial Expression Estimation in Egocentric Video

Yisi Zhang[1], Emma J. Jackson[2], Nicoleta Gavrila[2], Quoc C. Vuong[3], Bing Yuan[1] & Elena Geangu (elena.geangu@york.ac.uk)[2]

[1]Department of Psychological and Cognitive Sciences, Tsinghua University
[2]Department of Psychology, University of York
[3]School of Psychology & Biosciences Institute, Newcastle University

Abstract

Wearable head-mounted cameras offer a window into infant's social world, yet manual coding remains a computational bottleneck. To address this limitation, we evaluate a scalable, open-source pipeline for automated facial expression estimation using the EfficientNet-B0 architecture on naturalistic, egocentric videos. Despite non-canonical viewpoints, poor lighting and motion artifacts, the model achieves performance ($F1 = 0.543$) comparable to state-of-the-art benchmarks. We demonstrate representational alignment ($r = 0.76$) between model confusion patterns and human perception, suggesting that the model reflects the adult social-perceptual structure. While categorical accuracy varies, the model's distributional probabilities capture the graded nature of affect. We also identify a critical divergence in neutral expression processing, where model and human ambiguity dissociate ($r = -0.26$). These findings support automated facial expression estimation as a viable approach for quantifying infants' natural social input, while highlighting category-specific limitations. Together they represent a step toward a "big data" science of early social development.

Keywords: facial emotion expressions; infant; automated face detection; automated facial emotion expression estimation; real world big data

Introduction

The ability to infer affective states from facial expressions is a cornerstone of human sociality, underpinning many processes and phenomena, such as empathy and social learning (e.g., Eggleston et al., 2021; Denham et al., 2011; Geangu, 2009; Geangu, 2015). Research indicates that the development of this ability is protracted, emerging in infancy and continuing throughout childhood (e.g., Geangu et al., 2016; Herba and Phillips, 2004; López-Morales et al., 2025; Romani-Sponchiado et al., 2022; Ruba and Repacholi, 2020). However, while the points of inflection of this developmental trajectory are increasingly well-documented, the underlying mechanisms, specifically how infants transition from perceiving visual input to constructing rich emotional representations, remain poorly understood (e.g., Hoemann et al., 2020).

Traditionally, developmental research has relied on laboratory-based paradigms using predominantly 'canonical' facial displays: posed, frontal, and high-intensity expressions. There is a growing consensus that these stimuli lack the ecological validity necessary to explain how infants construct emotional representations from the natural statistics of social input in their daily lives (e.g. Hoemann et al., 2020; Ruba and Repacholi, 2020). In the natural environment, emotions are dynamic, multidimensional, and contextually modulated, reflecting communicative purposes and social norms (e.g., Crivelli and Fridlund, 2019; Jack et al., 2014; Reschke et al., 2018). Furthermore, infants are active perceivers in these social transactions (e.g., Denham et al., 2004; Özden et al., 2025; Smith et al., 2018), and therefore likely learn the meaning of facial displays from signals that are more subtle, fleeting, and contextualized than those traditionally studied in the lab (e.g., Dawel et al., 2023; LoBue et al., 2025; Steward et al., 2025).

To capture these natural statistics, the field has transitioned toward ecological sensing using head-mounted cameras (HMCs) to record the child's egocentric view (e.g., Geangu et al., 2023; Long et al., 2022; Smith et al., 2018). Such large ecological datasets provide a high-fidelity account of the frequency, variety, and context of the emotional signals infants and young children actually encounter in daily life (e.g., Long et al., 2022; Jayaraman and Smith, 2019). However, ecological sensing creates a substantial computational bottleneck. Manual coding is unfeasible at this scale (Long et al., 2022), and commercially available automated tools - GUI-dependent, proprietary, and optimized for high-fidelity lab data - drop to near-chance accuracy on spontaneous expressions and the non-canonical, off-angle views characteristic of infants' 'in-the-wild' egocentric footage (e.g., Dupré et al., 2020; Long et al., 2022; Jayaraman and Smith, 2019; Yurkovic-Harding and Bradshaw, 2024). These tools also lack the throughput required for large-scale longitudinal egocentric corpora (Long et al., 2022; Yurkovic-Harding and Bradshaw, 2024) and are difficult to integrate into scalable pipelines.

There is, therefore, an urgent need for open-source 'lightweight' yet robust deep learning architectures that provide a favorable trade-off between inference speed and estimation accuracy, and are good scalable solutions. In the present study, we evaluate an open-source pipeline, integrating RetinaFace (ResNet-50 backbone, Deng et al., 2020) for face detection with EfficientNet-B0 (Savchenko et al., 2022; Savchenko, 2025; Tan and Le, 2019) for emotion expression estimation, on a curated dataset of naturalistic egocentric video clips (EgoWild-Affect). Moving beyond accuracy-based evaluation, we leverage multi-rater labeling to examine model–human representational alignment, treating variability across observers as informative rather than noise (Cabitza et al., 2022; Cabitza et al., 2023). By validating this scalable approach against naïve human observers, we aim to facilitate a new era of developmental affective science.

Methods

Participants

Naïve adult raters (N = 149; 115 female, 25 male, 9 other; predominantly Caucasian/White Western background) were recruited from a UK-based university. Most participants were undergraduate Psychology students. Each provided informed consent prior to the task.

The EgoWild-Affect Dataset

We curated a benchmark dataset, the EgoWild-Affect (N = 316 video clips), to reflect the visual complexity of everyday social interactions involving infants, children, and adults. Clips were extracted from a 600-hours long corpus of HMC footage recorded in home (N = 251) and naturalistic laboratory (N = 65) settings, allowing for natural expressions in different contexts. The dataset captures 57 unique individuals, including infants/toddlers (N = 43, 25 female), children (N = 162, 95 female), and adults (N = 111, 95 female).

The clips ($M = 1.80s$, $Min = 0.27s$, $Max = 5s$) were selected to contain a single face displaying one of the six basic expressions (Anger, Disgust, Fear, Happiness, Sadness, Surprise) or an emotionally neutral expression. Based on majority-vote labels, the dataset comprised 54 anger, 76 happiness, 35 sadness, 29 disgust, 54 surprise, 12 fear, and 56 neutral video clips. The footage was recorded using miniature CMOS sensors recording at approximately 30 fps (1920 x 1080 pixels). These devices were embedded in custom headbands designed to capture the wearer's egocentric visual field.

Human rating protocol

Participants were semi-randomly assigned to rate approximately 80 clips each (37.3 ± 0.4 ratings/clip). Clips were presented muted at full- and half-speed. For each video clip, participants reported: (1) the predominant emotion expressed by the face (7-alternative forced choice: Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral); (2) the confidence in this choice (1-5 Likert scale, 1 - "Not at all confident", 5 = "Very confident"; and (3) the intensity with which the emotion was displayed (1-5 Likert scale, 1 - "Very weakly", 5 = "Very strongly", and N/A for neutral).

Automatic Facial Emotion Expression Estimation

We implemented a two-stage pipeline optimized for both robustness and computational scalability, essential for high-throughput processing of ecological data (Fig. 1).

Stage 1: Face Detection and Spatial Normalization Frames I_t were processed independently using RetinaFace (Res-Net-50 backbone; Deng et al., 2020). This model was selected for its ability to identify partial and heavily rotated faces, which are ubiquitous in infant- and toddler-view footage. It outputs three coupled variables for every candidate face: (1) Bounding box (b) defining the face region; (2) Confidence score (s): a scalar probability $p \in [0, 1]$ indicating the likelihood of a face; (3) Facial landmarks (l): a set of five coordinate pairs of the left/right eyes, nose tip, and left/right mouth corners. Detections were retained only when meeting a spatial confidence threshold $s > 0.7$. Surviving detections underwent a similarity transformation using the facial landmarks to map faces onto a 112×112 pixel reference frame, correcting for in-plane rotation and scale variations.

Stage 2: Facial Expression Classification and Temporal Smoothing Aligned RGB tensors ($112 \times 112 \times 3$) served as input to the EfficientNet-B0 architecture (Tan and Le, 2019). The B0 variant was prioritized for its superior trade-off between accuracy and inference speed (Savchenko, 2022). The classification head was modified to output a raw logit vector $z \in \mathbb{R}^7$, with Softmax activation generating a normalized probability distribution P across the seven categories (Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral), such that $\sum P = 1$. To mitigate frame-level jitter common in mobile sensing, probabilities were smoothed using a 10-frame moving average. To ensure that classification metrics reflected the performance of the expression classification stage rather than upstream face detection failures, we performed supervised verification of detected faces, retaining only true-positive detections for the present analyses. In deployed use, supervised verification operates as a human-in-the-loop quality control step on a subset of detections - a standard component of scalable computer vision pipelines for naturalistic data, where fully autonomous processing remains unfeasible (e.g., Mathis et al., 2018; Pereira et al., 2022).

Data analysis

Human perception analysis To quantify perceived confidence and intensity, we took the mean of the highest rated category. The normalized entropy was defined as $H(\mathbf{p}) = \frac{-\sum_{k=1}^{K} p_k \log p_k}{\log K}$, where $\mathbf{p} = (p_1, \ldots, p_K)$ denotes the categorical proportion. $H = 0$ indicates complete certainty, while $H = 1$ indicates maximal ambiguity. To characterize the structure of facial expression perception, we performed hierarchical clustering on the similarity between expression categories. Pairwise similarity between categories was quantified using Pearson correlation between these probability profiles. The resulting similarity matrix was converted to a dissimilarity matrix using $d = \sqrt{2(1 - r)}$, where r denotes the correlation coefficient. Hierarchical clustering was then performed on the dissimilarity matrix using average linkage, and the resulting dendrogram was used to order categories in the corresponding correlation matrices.

Model performance metrics Overall (top-1) accuracy was defined as the proportion of video clips for which the predicted category with the highest softmax probability matched the highest human voted label. Per-class accuracy was computed as the proportion of correctly classified clips within each expression category. Per-class AUC was computed us-

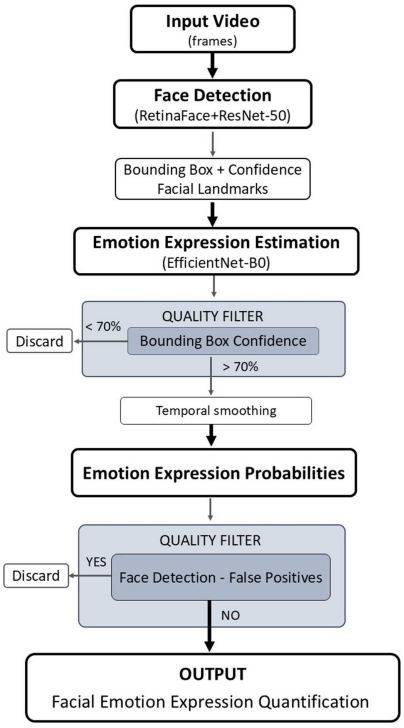

Figure 1: Schematic illustration of the automatic face detection and emotion expression estimation procedure.

ing a one-vs-rest approach, in which the softmax probability for a given expression category was contrasted against all other categories, and the area under the resulting ROC curve was calculated. F1 scores were computed for each category from the confusion matrix derived from top-1 predictions, with precision (P) defined as the proportion of predicted instances of a category that were correct and recall (R) as the proportion of true instances of that category that were correctly identified. F1 score was defined as $F1 = \frac{2PR}{P+R}$.

Results

Human agreement and confidence in facial expression classification

We first examined the reliability of human judgments. Consistent with the inherently ambiguous nature of spontaneous facial expressions in real-world contexts (e.g., Dawel et al., 2023; Steward et al., 2025), human raters did not uniformly agree on expression labels. The median per-clip agreement was 73%, indicating that on average approximately three-quarters of raters selected the same category for a given clip. Only 14.6% of the clips reached near-consensus levels (>95% agreement), whereas the lowest agreement observed was 24%, reflecting substantial ambiguity for some stimuli (Fig. 2A). These results indicate pronounced variability in human expression classification even under forced-choice conditions.

To assess whether inter-rater disagreement was related to subjective certainty, we examined the relationship between the proportion of majority votes and the mean confidence reported for the majority category. We observed a strong positive correlation (Spearman's $\rho = 0.79$, $p = 5.3 \times 10^{-69}$), indicating that higher inter-rater agreement was associated with greater self-reported confidence. We further examined whether perceived expression intensity was related to agreement across raters. Excluding neutral expressions, intensity ratings were positively correlated with agreement ($\rho = 0.62$, $p = 1.8 \times 10^{-28}$; Fig. 2B), suggesting that expressions perceived as more intense were also more consistently classified.

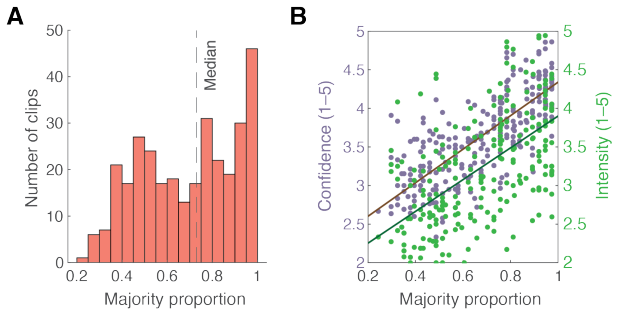

Figure 2: Human annotation agreement. (A) Histogram of the human rating proportion of the highest voted category. Dashed line indicates median. (B) Scatter plots of mean confidence (left) and intensity (right) of the highest voted category vs. the majority proportion.

Structure of human facial expression perception

We next examined whether human judgments exhibit systematic structure across expression categories. To quantify the clarity of each category, we compared the proportion of majority votes across expressions. Happiness elicited significantly greater agreement (84%) compared to others ($F(6, 309) = 11.4$, $p = 1.6 \times 10^{-11}$; one-way ANOVA with Tukey post-hoc tests; Fig. 3A). Consistent with this pattern, the entropy of raters' vote distributions (normalized) was significantly lower for happiness than for other categories ($F(6, 309) = 16.2$, $p = 3.4 \times 10^{-16}$; Fig. 3B), indicating greater consistency in its perception.

Confidence ratings mirrored these effects: happiness, sadness, and surprise were associated with higher reported confidence relative to other expressions ($F(6, 309) = 14.5$, $p = 1.3 \times 10^{-14}$; Fig. 3C). Differences in perceived intensity across expression categories were modest but statistically significant ($F(5, 254) = 2.49$, $p = 0.032$; Fig. 3D).

Finally, hierarchical clustering based on between-category correlations in human response profiles revealed a structured organization of facial expression perception (Fig. 3E). Disgust and anger clustered closely together, as did fear and surprise, whereas happiness formed a distinct and isolated cluster, reflecting its unique perceptual signature. Together, these analyses indicate that human facial expression recognition is characterized by systematic structure shaped by agreement, confidence, and perceived intensity, rather than being solely

determined by categorical labels.

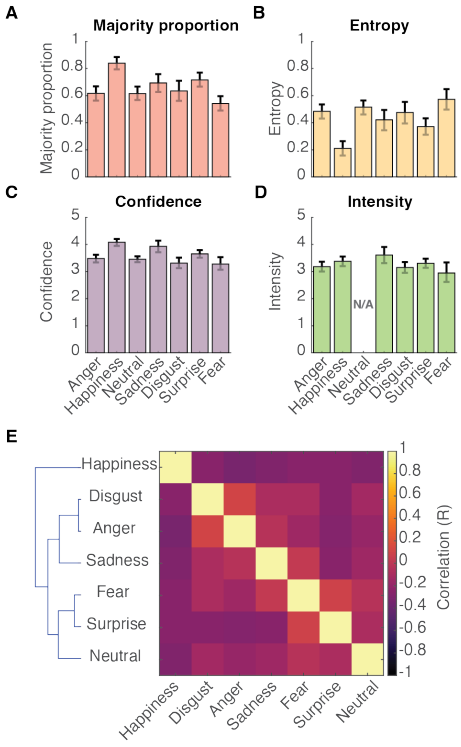

Figure 3: Human perception features. (A-D) Per-class majority proportion, entropy, confidence, and intensity. (E) Correlation matrix ordered by hierarchical clustering dendrogram. Error bars are bootstrap 95% confidence intervals.

Model performance

Overall, model predictions showed substantial agreement with the human majority-vote labels, achieving a top-1 accuracy of 61.1%, an averaged F1 score of 0.543 (Table 2), and a macro-averaged AUC of 0.87, indicating strong discriminability between expression categories. Performance varied systematically across expressions. Happiness (AUC = 0.96), surprise (AUC = 0.93), and sadness (AUC = 0.90) exhibited the highest ROC AUC values, reflecting robust separability from other categories (Fig. 4A).

Per-class top-1 accuracy averaged 0.55 ± 0.21, whereas top-2 accuracy reached 0.80 ± 0.10 (mean ± sd), indicating substantial gains when graded probabilistic information was taken into account. Classification performance differed markedly across expression categories (likelihood-ratio test, $\chi^2 = 37.6$, $p = 1.33 \times 10^{-6}$; Fig. 4B). Fear exhibited the lowest top-1 accuracy (0.25), yet showed a great improvement in top-2 accuracy (0.83), suggesting that although the model rarely selected fear as the single most likely category, it often assigned high probability mass to it among the top candidates. In contrast, happiness, sadness, and surprise achieved significantly higher top-1 accuracy than disgust, neutral, and fear (pairwise comparisons with FDR correction, adjusted $q < 0.05$; Fig. 4B). Notably, this performance hierarchy closely mirrored the pattern of human inter-rater agreement, indicating that categories that were perceptually clearer to humans were also more reliably classified by the model.

Finally, we examined how model accuracy related to human perceptual measures. Accuracy decreased systematically with increasing human ambiguity, quantified by entropy of the vote distribution (Fig. 4C). Similarly, clips associated with lower rated confidence (Fig. 4D) and lower perceived expression intensity (Fig. 4E) were more likely to be misclassified.

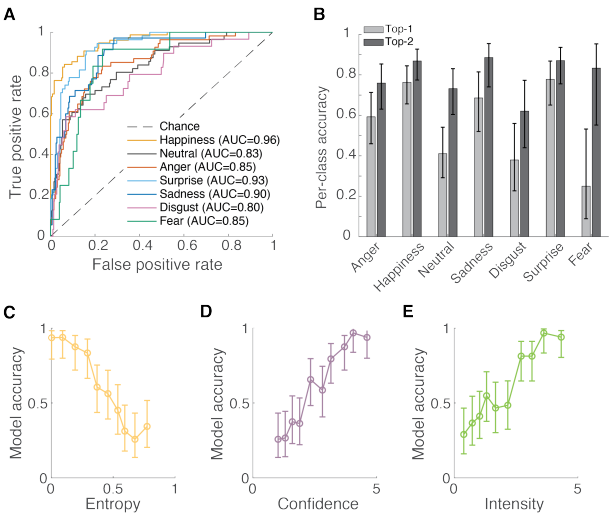

Figure 4: Model performance. (A) Per class ROC and AUC. (B) Top-1 and top-2 per class accuracy. (C-E) Model accuracy vs. entropy, confidence, and intensity. Error bars are 95% confidence intervals.

Human–model alignment in facial expression perception structure

To assess whether the model exhibited a facial-expression recognition structure comparable to human perception, we compared two confusion matrices. The first was constructed from human majority labels against the expected category probabilities derived from raters' vote proportions, and the second from the same majority labels against the model's top-1 predictions (Fig. 5A). In both cases, rows were normalized to sum to one, such that each entry reflected the conditional probability of assigning a given expression category, given the true category.

Across all off-diagonal entries, model confusion strengths were systematically higher than those of humans ($p = 1.0 \times 10^{-9}$, paired t-test), indicating greater overall uncertainty in the model's categorical assignments. Despite this difference in magnitude, human and model confusion patterns were strongly correlated (Pearson's $r = 0.76$, $p = 4.1 \times 10^{-9}$; Fig. 5B), suggesting that the model tends to confuse expression categories in a manner consistent with human perceptual confusions rather than arbitrarily.

To further examine the structure of the model's representations, we computed a correlation matrix of the model's pre-

dicted probability profiles across expression categories and applied hierarchical clustering. Consistent with the structure observed in human annotations, the model exhibited high similarity between fear and surprise, as well as between disgust and anger (Fig. 5C). However, the model additionally showed moderate similarity between neutral and sadness, a relationship that was not evident in human perceptual data. Together, these findings indicate that while the model largely approximates the structure of human facial-expression perception, it also introduces category-specific deviations.

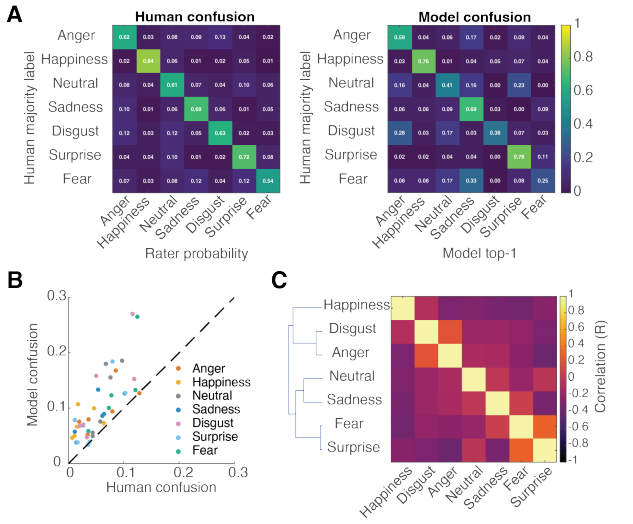

Figure 5: Alignment of confusion structure. (A) Row-normalized confusion matrix of human ratings (left) and model prediction (right). (B) Scatter plot of the off-diagonal elements between model and human confusion. (C) Correlation matrix ordered by hierarchical clustering dendrogram.

Finally, we examined whether expressions that were more ambiguous to humans were also more ambiguous to the model. Across all clips, the entropy of the model's softmax output was positively correlated with the entropy of human ratings ($r = 0.41$, $p = 3.4 \times 10^{-14}$), indicating shared sensitivity to perceptual ambiguity. However, this relationship differed across expression categories. In particular, neutral expressions showed a significantly lower correlation compared to other categories (interaction model with FDR-corrected post-hoc tests, $p = 3.5 \times 10^{-4}$; Fig. 6), and in fact exhibited a negative correlation ($r = -0.26$, $p = 0.049$).

Discussion

The present study evaluates a scalable open-source pipeline for automated facial expression estimation in naturalistic egocentric video, and examines the extent to which its outputs align with human perceptual judgments. While laboratory-based models often falter on spontaneous expressions, particularly under the non-canonical viewpoints and motion artifacts of egocentric footage, the EfficientNet-B0 architecture extracts meaningful facial affective signals from raw video data at a level that approximates human perceptual structure.

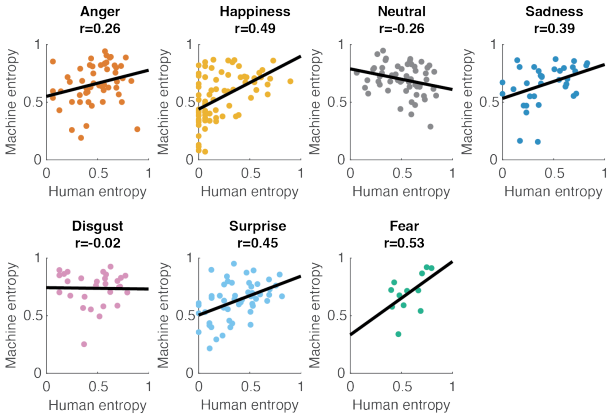

Figure 6: Alignment of perception ambiguity for each expression category.

Table 1: Comparison with other datasets and models.

Dataset	Method	Acc. (%)	Avg. F1
AFEW	EfficientNet-B0 (Savchenko et al., 2022)	59.3	-
	EfficientNet-B2	59.0	-
	MobileNet-v1	55.4	-
Aff-Wild2 (ABAW-3)	EfficientNet-B0 (Savchenko, 2022)	50.0	0.381
	EfficientNet-B2	43.5	0.317
	MobileNet	46.0	0.338
Aff-Wild2 (ABAW-6)	Multi-modal (W. Zhang et al., 2024)	-	0.501
Aff-Wild2 (ABAW-8)	wav2vec 2.0 + Emoti-EffNet	55.3	0.446
Our data	EfficientNet-B0	**61.1**	**0.543**

Previous work has shown that EfficientNet-B0 outperforms other lightweight architectures for facial expression recognition in unconstrained settings (Savchenko et al., 2022). We extend these findings by demonstrating that the model generalizes well on egocentric developmental data, achieving comparable performance with SOTA (Table 1; Kollias et al., 2024, 2025). At the per-class level, performance was on par with, and in many cases exceeded, results obtained on the Aff-Wild2 dataset using EfficientNet-B0 and related state-of-the-art architectures for nearly all expression categories (Table 2; Savchenko, 2022; W. Zhang et al., 2024). We show that although the egocentric home videos frequently contained extreme angles and partial occlusions, our model consistently detected facial regions associated with expression perception (Fig. 7). Taken together, our pipeline is an efficient adaptation to the more extreme viewpoints and potentially variable cases characteristic of infant perspective footage.

A central finding is the high degree of representational alignment between the model and human perception. This is evident both in terms of agreement and confusion, particularly for facial displays with morphologically similar signals (e.g., Anger and Disgust; Fear and Surprise). A complementary view of these findings is that aggregate accuracy, per-class accuracy, and confusion structure are not independent metrics — they are three expressions of a single prop-

Table 2: Per class F1 score comparison with Aff-Wild2 predicted by SOTA models.

Expression	Savchenko 2022	Zhang 2024	Our data
Anger	0.151	**0.739**	0.587
Happiness	0.477	0.591	**0.811**
Neutral	0.609	**0.702**	0.484
Sadness	0.461	**0.664**	0.552
Disgust	**0.516**	0.503	0.478
Surprise	0.303	0.365	**0.689**
Fear	0.016	**0.218**	0.200

erty: the model is calibrated to the same perceptual structure as human raters, succeeding where humans agree and struggling where humans disagree. Categories with high human consensus (happiness, surprise, sadness) are classified at or above SOTA benchmarks (Table 2), while lower accuracy for fear, neutral, and disgust mirrors the categories where human raters themselves disagree most (Figure 3A). This pattern is consistent with the well-documented ambiguity of these categories in spontaneous, naturalistic expressions, where signals are subtle, mixed, and contextually modulated (e.g., Dawel et al., 2023; Jack et al., 2016; Steward et al., 2025). Under these conditions, a classifier that substantially exceeded human reliability would be a less faithful proxy for human social perception, not a better one. The aggregate accuracy of 61% therefore reflects a perceptual ceiling of the task more than a limit of model capability — though the precision of this estimate is itself bounded by label uncertainty on clips with low agreement by humans, an inherent feature of naturalistic affective stimuli that should be addressed in future work through expert coding or larger samples of naïve human raters.

Our findings also advocate for a shift toward using distributional probabilities of facial expressive signal as a more accurate representation of the infant's visual input. The model's softmax output provides a granular measure of these fleeting and mixed signals. For instance, the leap from 25% top-1 accuracy to 83% top-2 accuracy for Fear suggests that the model identifies relevant affective information even when it is not the dominant signal. This would be missed if we were to simply rely on dominant categorical labels. Utilizing the full probability distributions may allow researchers to better characterize the more subtle, varied, and contextualized facial affective displays from which infants are learning.

Notably, there are also points of representational divergence between the model and human observers. This was mostly evident for neutral faces, where increased human confidence and reduced ambiguity corresponded to greater model uncertainty. This dissociation likely reflects two compounding factors. First, the training data (AffectNet, AFEW) consists predominantly of posed or acted expressions, where 'neutral' is a visually prototypical category; in naturalistic footage, neutral faces may be easily confused with subtle expressions, a distribution the model has not been optimized for. Second, human observers are known to over-attribute affect, traits, and intentions to neutral faces, which can produce high subjective confidence even when the underlying signal is genuinely ambiguous (Said et al., 2009). Researchers should therefore exercise caution when interpreting model-estimated "neutral" facial displays, recognizing that this is the category where model and human judgments diverge in systematic ways.

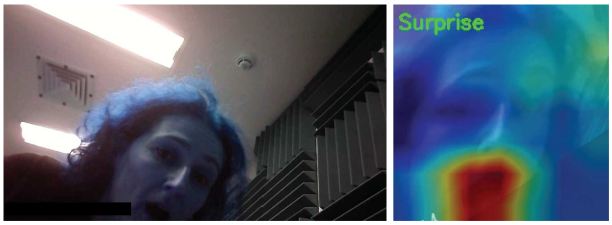

Figure 7: Visualizing EfficientNet-B0 attention with Grad-CAM (Selvaraju et al., 2016). The heatmap overlays indicate regions contributing most to the model's prediction (warmer colors - higher importance), showing that the model relies on discriminative facial features (e.g., mouth curvature, eye region) rather than background to classify expressions.

Future directions As part of a larger effort to develop scalable methodologies for studying social development with high ecological validity from the first days of life (e.g., Geangu et al., 2023; Y. Zhang et al., 2025; Mason et al., 2024), future research could leverage emerging corpora of egocentric audio-video footage to increase the number of video clips included in the EgoWild-Affect dataset, and perform domain-specific fine-tuning. Training models with larger infant-perspective datasets is expected to increase robustness and generalization to address some of the present limitations. One of these is the class imbalance, particularly the small number of fear video clips, due to the difficulty of finding spontaneously elicited fearful facial expressions (see also Dupré et al., 2020). A further priority is extending cultural representation beyond the predominantly Western Caucasian sample of video clips and raters included here.

Beyond scaling, the field should also transition to multimodal emotion expression recognition. Early social communication is inherently multimodal (e.g., Crespo-Llado et al., 2018; Geangu et al., 2016; Geangu and Vuong, 2020; Geangu and Vuong, 2023; Ke et al., 2022; Riviere et al., 2026; Vuong and Geangu, 2023), thus integrating facial cues with automated vocal and postural estimation is necessary. Furthermore, it would also be important to consider the wider context in which affective displays are perceived. People's speech and actions, as well as infants' own behaviour and physiological responses, have all been hypothesized to play a significant role in learning the communicative meaning of affective displays (e.g., Amso and Kirkham, 2021; Addabbo and Turati, 2020; Addabbo et al., 2026; Hoemann et al., 2020; Poulin-Dubois et al., 2018). Ultimately, this high-throughput approach will provide richer insights into the cascading effects of change in other domains on the development of emotion understanding (e.g., LoBue et al., 2025; Oakes, 2017).

Acknowledgments

We thank all participants who made this study possible. We would also like to thank William AP Smith for his technical support. This work was supported in part by the Wellcome Trust Institutional Strategic Support Fund (ISSF) and the Engineering and Physical Sciences Research Council (EPSRC) Impact Acceleration Account awarded to the University of York (allocated to Elena Geangu), by the Wellcome Leap 1kD Program (funds awarded to Elena Geangu - Lead PI), and by a Google DeepMind award to Elena Geangu.

References

Addabbo, M., Mermier, J., Rutkowska, J., Meyer, M., Hunnius, S., Turati, C., Bulf, H., et al. (2026). The infant brain combines emotional information from faces and action kinematics. *Journal of Experimental Child Psychology*, *263*, 1–9.

Addabbo, M., & Turati, C. (2020). Binding actions and emotions in the infant's brain. *Social neuroscience*, *15*(4), 470–476.

Amso, D., & Kirkham, N. (2021). A multiple-memory systems framework for examining attention and memory interactions in infancy. *Child Development Perspectives*, *15*(2), 132–138.

Cabitza, F., Campagner, A., & Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(6), 6860–6868.

Cabitza, F., Campagner, A., & Mattioli, M. (2022). The unbearable (technical) unreliability of automated facial emotion recognition. *Big data & society*, *9*(2), 20539517221129549.

Crespo-Llado, M. M., Vanderwert, R. E., & Geangu, E. (2018). Individual differences in infants' neural responses to their peers' cry and laughter. *Biological psychology*, *135*, 117–127.

Crivelli, C., & Fridlund, A. J. (2019). Inside-out: From basic emotions theory to the behavioral ecology view. *Journal of nonverbal behavior*, *43*(2), 161–194.

Dawel, A., Ashhurst, C., & Monaghan, C. (2023). A three-dimensional model of emotional display rules: Model invariance, external validity, and gender differences. *Emotion*, *23*(5), 1410.

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5203–5212.

Denham, S., Warren, H., Von Salisch, M., Benga, O., Chin, J., & Geangu, E. (2011). Emotions and social development in childhood. In *The wiley-blackwell handbook of childhood social development: Second edition* (pp. 413–433).

Denham, S., Caal, S., Bassett, H. H., Benga, O., & Geangu, E. (2004). Listening to parents: Cultural variations in the meaning of emotions and emotion socialization. *Cogniţie Creier Comportament*.

Dupré, D., Krumhuber, E. G., Küster, D., & McKeown, G. J. (2020). A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *Plos one*, *15*(4), e0231968.

Eggleston, A., Geangu, E., Tipper, S. P., Cook, R., & Over, H. (2021). Young children learn first impressions of faces through social referencing. *Scientific Reports*, *11*(1), 14744.

Geangu, E. (2009). Empathy development-insights from early years: Introduction to the special issue. *Cognition, brain, behavior*, *13*(4), 363–366.

Geangu, E. (2015). Development of empathy during early childhood across cultures. In *International encyclopedia of the social and behavioral sciences* (pp. 549–553). Elsevier.

Geangu, E., Quadrelli, E., Conte, S., Croci, E., & Turati, C. (2016). Three-year-olds' rapid facial electromyographic responses to emotional facial expressions and body postures. *Journal of experimental child psychology*, *144*, 1–14.

Geangu, E., Smith, W. A., Mason, H. T., Martinez-Cedillo, A. P., Hunter, D., Knight, M. I., Liang, H., del Carmen Garcia de Soria Bazan, M., Tse, Z. T. H., Rowland, T., et al. (2023). Egoactive: Integrated wireless wearable sensors for capturing infant egocentric auditory–visual statistics and autonomic nervous system function 'in the wild'. *Sensors*, *23*(18), 7930.

Geangu, E., & Vuong, Q. C. (2020). Look up to the body: An eye-tracking investigation of 7-months-old infants' visual exploration of emotional body expressions. *Infant Behavior and Development*, *60*, 101473.

Geangu, E., & Vuong, Q. C. (2023). Seven-months-old infants show increased arousal to static emotion body expressions: Evidence from pupil dilation. *Infancy*, *28*(4), 820–835.

Herba, C., & Phillips, M. (2004). Development of facial expression recognition from childhood to adolescence: Behavioural and neurological perspectives. *Journal of Child Psychology and Psychiatry*, *45*(7), 1185–1198.

Hoemann, K., Wu, R., LoBue, V., Oakes, L. M., Xu, F., & Barrett, L. F. (2020). Developing an understanding of emotion categories: Lessons from objects. *Trends in cognitive sciences*, *24*(1), 39–51.

Jack, R. E., Garrod, O. G., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, *24*(2), 187–192.

Jack, R. E., Sun, W., Delis, I., Garrod, O. G., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General*, *145*(6), 708.

Jayaraman, S., & Smith, L. B. (2019). Faces in early visual environments are persistent not just frequent. *Vision research*, *157*, 213–221.

Ke, H., Vuong, Q. C., & Geangu, E. (2022). Three-and six-year-old children are sensitive to natural body expressions of emotion: An event-related potential emotional priming study. *Journal of Experimental Child Psychology*, *224*, 105497.

Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Kotsia, I., Baird, A., Gagne, C., Shao, C., & Hu, G. (2024). The 6th affective behavior analysis in-the-wild (abaw) competition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4587–4598.

Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Kotsia, I., Granger, E., Pedersoli, M., Bacon, S., Baird, A., Gagne, C., et al. (2025). Advancements in affective and behavior analysis: The 8th abaw workshop and competition. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5572–5583.

LoBue, V., Casasola, M., & Oakes, L. M. (2025). Emotion understanding in infants and young children: How input shapes emotional development. *Advances in child development and behavior*, *69*, 69–98.

Long, B. L., Kachergis, G., Agrawal, K., & Frank, M. C. (2022). A longitudinal analysis of the social information in infants' naturalistic visual experience using automated detections. *Developmental Psychology*, *58*(12), 2211.

López-Morales, H., Zabala, M., Agulla, L., Aguilar, M., Sosa, J., Vivas, L., & López, M. (2025). Accuracy and speed in facial emotion recognition in children, adolescents, and adults. *Current Psychology*, *44*(6), 4356–4370.

Mason, H. T., Martinez-Cedillo, A. P., Vuong, Q. C., Garcia-de-Soria, M. C., Smith, S., Geangu, E., & Knight, M. I. (2024). A complete pipeline for heart rate extraction from infant ecgs. *Signals*, *5*(1), 118–146.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, *21*(9), 1281–1289.

Oakes, L. M. (2017). Plasticity may change inputs as well as processes, structures, and responses. *Cognitive development*, *42*, 4–14.

Özden, Z. B., Walle, E. A., & Reschke, P. J. (2025). Infant-centered behavioral response patterns to discrete emotions. *Developmental Psychology*.

Pereira, T. D., Tabris, N., Matsliah, A., Turner, D. M., Li, J., Ravindranath, S., Papadoyannis, E. S., Normand, E., Deutsch, D. S., Wang, Z. Y., et al. (2022). Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, *19*(4), 486–495.

Poulin-Dubois, D., Hastings, P. D., Chiarella, S. S., Geangu, E., Hauf, P., Ruel, A., & Johnson, A. (2018). The eyes know it: Toddlers' visual scanning of sad faces is predicted by their theory of mind skills. *PLOS one*, *13*(12), e0208524.

Reschke, P. J., Knothe, J. M., Lopez, L. D., & Walle, E. A. (2018). Putting "context" in context: The effects of body posture and emotion scene on adult categorizations of disgust facial expressions. *Emotion*, *18*(1), 153.

Riviere, E., Courbois, Y., & Gentaz, E. (2026). The developmental changes in emotion recognition from human biological motion by children aged from 4 to 12 years [Advance online publication]. *Emotion*. https://doi.org/10.1037/emo0001626

Romani-Sponchiado, A., Maia, C. P., Torres, C. N., Tavares, I., & Arteche, A. X. (2022). Emotional face expressions recognition in childhood: Developmental markers, age and sex effect. *Cognitive Processing*, *23*(3), 467–477.

Ruba, A. L., & Repacholi, B. M. (2020). Do preverbal infants understand discrete facial expressions of emotion? *Emotion Review*, *12*(4), 235–250.

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, *9*(2), 260–264. https://doi.org/10.1037/a0014681

Savchenko, A. V. (2022). Video-based frame-level facial analysis of affective behavior on mobile devices using efficientnets. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2359–2366.

Savchenko, A. V. (2025). Hsemotion team at abaw-8 competition: Audiovisual ambivalence/hesitancy, emotional mimicry intensity and facial expression recognition. *arXiv preprint arXiv:2503.10399*.

Savchenko, A. V., Savchenko, L. V., & Makarov, I. (2022). Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, *13*(4), 2132–2143.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in cognitive sciences*, *22*(4), 325–336.

Steward, B. A., Mewton, P., Palermo, R., & Dawel, A. (2025). Interactions between faces and visual context in emotion perception: A meta-analysis. *Psychonomic Bulletin & Review*, 1–17.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning*, 6105–6114.

Vuong, Q. C., & Geangu, E. (2023). The development of emotion processing of body expressions from infancy to early childhood: A meta-analysis. *Frontiers in Cognition*, *2*, 1155031.

Yurkovic-Harding, J., & Bradshaw, J. (2024). Automated detection of faces in infant and parent first-person views during play. *2024 IEEE International Conference on Development and Learning (ICDL)*, 1–6.

Zhang, W., Qiu, F., Liu, C., Li, L., Du, H., Guo, T., & Yu, X. (2024). An effective ensemble learning framework for affective behaviour analysis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4761–4772.

Zhang, Y., Martinez-Cedillo, A. P., Mason, H. T., Vuong, Q. C., Garcia-de-Soria, M. C., Mullineaux, D., Knight, M. I., & Geangu, E. (2025). An automatic sustained attention prediction (asap) method for infants and toddlers using wearable device signals. *Scientific Reports*, *15*(1), 13298.