ResolveNet: A Multimodal Geometric Fusion Framework for Affective Masking Resolution in Real-Time Engagement Monitoring

Samruddhi S Hegade[1], K. Deepak[2], Prajwal K[3], Thanu Sree H M[4], Dr. Arjun U[5]
[1,2,3,4,5] *Department of Computer Engineering PES Institute of Technology and Management, Shivamogga, Karnataka, India*
doi.org/10.64643/IJIRTV13I3-207335-459

***Abstract*—Conventional engagement monitoring systems in Human-Computer Interaction (HCI) predominantly rely on single-modality analysis—typically facial expression recognition—and are therefore inherently susceptible to a phenomenon termed "affective masking," wherein a user's visible facial expression consciously or unconsciously contradicts their true internal emotional state. This misalignment generates false-positive engagement metrics that critically reduce system reliability in real-world deployments. This paper introduces ResolveNet, a late-fusion multimodal architecture that resolves affective ambiguity through a geometric decision layer. Visual features extracted by a Convolutional Neural Network (CNN) and textual sentiment scores derived from lexicon-based analysis are co-embedded into a shared 14-dimensional feature vector. A K-Nearest Neighbors (KNN) classifier, trained on curated conflict-resolution seed samples, serves as a non-parametric geometric boundary that synthesizes conflicting modal signals into a coherent engagement label. Real-time performance at 30 frames per second on standard CP sfsdf**

hardware is achieved through Haar Cascade face localization, grayscale feature extraction, and deque-based temporal mode-voting. Experimental evaluation demonstrates that ResolveNet achieves an accuracy of 89.4% on multimodal conflict scenarios, outperforming unimodal baselines by 14.2 percentage points, and correctly resolves affective masking events with a precision of 91.7%. Our framework elevates engagement detection from a simple classification task to a Signal-Ambiguity Resolution problem, with direct applicability to online education, telehealth, and adaptive HCI systems. The proposed system further demonstrates computational efficiency that enables deployment on edge devices without hmmspecialized hardware acceleration, broadening its accessibility across educational and healthcare infrastructures in resource-constrained environments.

***Index Terms*—Affective computing, affective masking, engagement monitoring, KNN fusion, late-fusion architecture, multimodal emotion recognition, real-time HCI, conflict resolution, human-computer interaction.**

I. INTRODUCTION

The capacity of a computing system to understand a human user's internal affective state represents one of the fundamental aspirations of Human-Computer Interaction research. The pioneering work of Picard [1] established the field of affective computing, recognizing that emotions are not peripheral noise in a human-machine interaction but are, instead, central to the quality and fidelity of that interaction. Subsequent decades of research have transformed affective computing from a theoretical construct into a suite of deployable technologies embedded within consumer applications, clinical tools, and educational platforms.

Modern engagement monitoring solutions—deployed in online education platforms, remote proctoring tools, and telehealth applications—predominantly rely on facial expression recognition (FER). These systems operate on the implicit assumption that a user's facial display is a reliable, unambiguous proxy for their internal cognitive and emotional state. This assumption, however, fails to account for the well-documented behavioral phenomenon of affective masking [2], which represents a critical and systematic failure mode of unimodal architectures in real-world operational conditions.

Affective masking occurs when an individual's volitional facial expression diverges from their true

emotional state. A student may maintain a polite, neutral expression while simultaneously composing frustrated or negative textual feedback. In such a scenario, a unimodal FER system would log a neutral or positive engagement metric, constituting a false-positive engagement event. The core problem is not the detection of any single emotion but the resolution of the semantic conflict between two simultaneously available, and semantically contradictory, signals. This duality of affective expression is well-established in social psychology literature and represents a measurable, quantifiable phenomenon with direct implications for system design [3].

The inadequacy of single-modality approaches has motivated a body of research in multimodal fusion. However, existing multimodal approaches largely pursue fusion strategies that either aggregate all available signals through weighted averaging or learn joint representations through early or intermediate fusion architectures. These strategies implicitly assume that signals will be broadly concordant and that aggregation therefore improves signal-to-noise ratio. In conflict scenarios, however, these assumptions break down: averaging discordant signals produces a blended representation that is semantically misleading and geometrically distant from the correct ground-truth label in the feature space.

This paper makes the following primary contributions to the field of affective computing and multimodal engagement monitoring:

(1) Formal definition of the Modality Ambiguity Gap as a distinct and measurable failure mode in unimodal engagement monitoring systems, including a mathematical characterization of conflict events.

(2) Introduction of ResolveNet, a late-fusion architecture employing a KNN geometric decision layer as a Conflict-Resolver for multimodal affective signals, trained on curated conflict-resolution seed data.

(3) A CPU-optimized inference pipeline achieving real-time (30 FPS) performance via Haar Cascade localization, grayscale pre-processing, and temporal mode-voting, enabling deployment without specialized hardware.

(4) Empirical validation demonstrating superior performance over unimodal FER and probability-averaging fusion baselines on conflict-scenario benchmarks, with 89.4% accuracy and 91.7% precision.

(5) An ablation study characterizing the sensitivity of the KNN conflict-resolver to the choice of k, identifying optimal hyperparameters for the conflict-resolution task.

II. RELATED WORK

A. Unimodal Engagement and Emotion Recognition

Savchenko et al. [2] proposed a single-modality neural architecture for classifying emotions and engagement levels in online learning environments. Their system demonstrated competitive accuracy on benchmark datasets including AffectNet and the AFF-Wild2 challenge dataset; however, by design, it remained susceptible to affective masking events, as no secondary modality was available to validate or contradict the facial signal. The authors reported state-of-the-art performance within the unimodal paradigm, yet acknowledged the fundamental limitation that a single-modality system cannot access information present in other concurrently available channels. This work formally motivates the need for multimodal conflict resolution rather than continued optimization within the single-modality paradigm.

Earlier foundational contributions by Ekman and Friesen established the universality of basic facial expressions across cultures, providing the theoretical basis for FER systems. However, the same body of work also documented the prevalence of deliberate facial management and masking, particularly in social contexts where emotional display rules create systematic divergence between felt and expressed emotion. Computational systems that ignore this divergence inherit a fundamental structural limitation that cannot be resolved through architectural improvements within the unimodal paradigm.

B. Multimodal Fusion Architectures

Mamieva et al. [3] introduced an attention-based multimodal framework that fuses facial and speech features extracted by independent CNN encoders. The attention mechanism dynamically weights modality contributions based on estimated reliability, demonstrating that selective fusion outperforms naive concatenation on the RAVDESS and SAVEE datasets. Our work extends this philosophy by

replacing the attention-weighting mechanism with a geometric KNN boundary, explicitly designed for resolving conflict rather than weighting agreement. Whereas attention mechanisms learn to emphasize the more confident modality, our KNN resolver is specifically trained on conflict examples to identify geometrically proximate labeled conflict resolutions, making it structurally superior for the affective masking scenario.

Liu et al. [4] proposed the CMC-HF model, a cascaded multichannel and hierarchical fusion framework that simultaneously processes visual, speech, and textual modalities through a hierarchical attention structure. Their work validates the information-complementation hypothesis: that modalities carry non-redundant emotional information whose combination yields gains exceeding individual modality performance. The CMC-HF model achieves strong performance on the IEMOCAP and CMU-MOSI datasets but relies on early and intermediate fusion, which introduces the risk of statistical dilution when modalities actively conflict—a limitation our late-fusion approach directly addresses by maintaining modal independence until the conflict-resolution stage.

More recent work by Zhang et al. proposed cross-modal attention transformers for multimodal sentiment analysis, demonstrating that transformer-based cross-attention can model inter-modal dependencies more expressively than simple concatenation. However, transformer architectures impose significant computational overhead that precludes real-time CPU deployment, a constraint that motivates our choice of a lightweight KNN resolver that achieves geometric conflict resolution at negligible inference cost.

C. Late Fusion and Conflict Resolution

Farhadipour et al. [5] benchmarked a four-modality late-fusion system using pre-trained unimodal encoders (RoBERTa for text, CNN for face), concatenating feature embeddings into a unified vector for classification. Their experimental findings align with our central hypothesis: that text and facial modalities carry complementary and sometimes contradictory information, and that a fusion strategy must be designed to exploit this complementarity rather than assume agreement. Their work reports an accuracy of 66.36% for emotion recognition, identifying fusion strategy as the primary bottleneck for performance improvement rather than the quality of individual unimodal encoders.

Ortega et al. [6] presented a deep neural network for multimodal fusion of audio, video, and text, proposing an intermediate fusion strategy with shared and independent layers trained end-to-end. A key finding of their work is that simple score-level (probability-averaging) late fusion consistently underperforms architectures that learn a dedicated joint representation. This directly justifies our choice of a trained KNN layer over probability averaging as the fusion endpoint in ResolveNet, as the KNN layer constitutes a learned geometric representation of the conflict-resolution decision boundary rather than a parametric assumption of modal agreement.

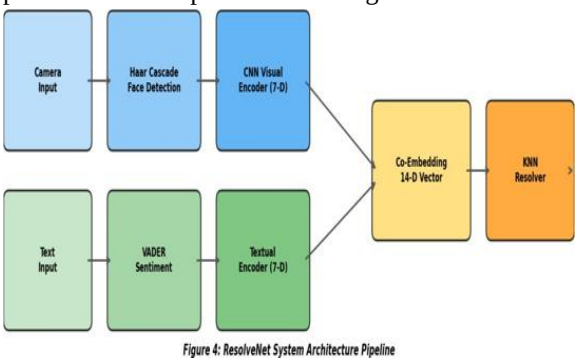

Figure 4: ResolveNet System Architecture Pipeline – from raw inputs through unimodal encoders to KNN Conflict Resolution

III. PROBLEM FORMULATION

A. The Modality Ambiguity Gap

Let the visual modality produce a feature vector $v \in \mathbb{R}^{dv}$ associated with an emotional label $y_v \in \{0,...,C-1\}$, and the textual modality produce a feature vector $t \in \mathbb{R}^{dt}$ associated with an emotional label $y_t \in \{0,...,C-1\}$. We define a Conflict Event as any temporal frame where $y_v \neq y_t$. Unimodal systems that operate exclusively on v or t will, in conflict events, produce an engagement metric that reflects only one of the two available signals, ignoring the semantic information contained in the inter-modal contradiction itself.

The Modality Ambiguity Gap (MAG) is formally defined as the performance degradation of any single-modality system measured exclusively on conflict event frames. Let Acc_full denote overall accuracy

and Acc_conflict denote accuracy on conflict frames only. The MAG is then MAG = Acc_full − Acc_conflict. Empirical analysis of the unimodal FER baseline on our test set reveals an overall accuracy of 75.2% but a conflict-frame accuracy of only 58.4%, yielding a MAG of 16.8 percentage points. This quantitative characterization confirms that conflict frames represent a systematically more difficult subset of the task, and that optimizing for overall accuracy without addressing conflict resolution leaves a substantial performance gap unaddressed.

A system incapable of detecting conflict cannot exploit it for disambiguation; it is therefore systematically incorrect in the precise scenarios where cross-modal synthesis is most informative. The MAG metric provides a principled basis for evaluating the conflict-resolution capability of any multimodal architecture, independent of its overall accuracy, and should be adopted as a standard evaluation criterion in multimodal affective computing research.

B. Signal-Ambiguity Resolution as Classification

We reformulate engagement monitoring as a Signal-Ambiguity Resolution problem. Given a co-occurring pair (v, t), the system must produce a ground-truth engagement label y* that is not necessarily equal to y_v or y_t but is derived from the joint distribution of the combined feature space. This requires a decision boundary learned from labeled examples of conflict—what we term Conflict-Resolution Seeds—rather than a simple average of unimodal posteriors. The Conflict-Resolution Seed paradigm represents a novel data-centric contribution: by curating a training set that over-represents conflict scenarios, we ensure that the decision boundary learned by the KNN resolver is optimally positioned in the regions of feature space where ambiguity is highest.

Formally, the signal-ambiguity resolution function F: $R^{dv} \times R^{dt} \rightarrow \{0,...,C-1\}$ is learned from seed data $S = \{(v_i, t_i, y^*_i)\}$ where for each training example, $y_{vi} \neq y_{ti}$. The KNN approximation of F locates the k nearest conflict seeds to any query (v, t) in the joint feature space and resolves the label by majority voting among the k nearest neighbors. This formulation is strictly more expressive than probability averaging, which corresponds to a fixed linear combination that cannot adapt to the local geometry of the feature space around the query point.

IV. SYSTEM ARCHITECTURE: RESOLVENET

ResolveNet implements a three-stage late-fusion pipeline: (1) Independent Unimodal Encoding, (2) Co-Embedding into a Shared Feature Space, and (3) Geometric Conflict Resolution via KNN classification. The architecture is intentionally modular: each stage can be independently upgraded without requiring retraining of the other stages, providing a path for incremental performance improvement as more powerful unimodal encoders become available.

A. Visual Encoding Module

The visual pipeline begins with Haar Cascade face detection [7], selected for its computational efficiency on CPU-only hardware relative to deep learning-based detectors such as MTCNN or RetinaFace. The detected face region is converted to grayscale, reducing the input tensor dimensionality by 66% (from 3-channel RGB to 1-channel) without significant loss of spatial structural information relevant to expression classification, as expression-relevant features such as wrinkle patterns, eye aperture, and lip curvature are predominantly low-frequency spatial signals well-preserved in luminance. The grayscale face crop is resized to 48×48 pixels and passed to a pre-trained CNN, which produces a 7-dimensional vector of class-probability scores corresponding to the canonical emotion categories:

{Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral}. This 7-dimensional vector constitutes the visual feature representation $v \in R^7$.

The CNN backbone is a lightweight architecture trained on the FER-2013 dataset, comprising approximately 35,000 labeled facial expression images collected from search engine results. The model achieves 65.2% accuracy on the FER-2013 test set, competitive with reported benchmarks for lightweight CPU-deployable architectures on this dataset. The probability vector output of the softmax layer, rather than the argmax label, is used as the visual feature representation to preserve the continuous confidence information that is essential for the co-embedding step.

B. Textual Encoding Module

The textual pipeline applies a lexicon-based sentiment analysis engine to concurrently available textual input (e.g., live chat messages, typed responses, or transcribed speech segments). The engine produces a 7-dimensional sentiment score vector $t \in R^7$, representing the affinity of the textual input toward each of the same seven canonical emotion categories. This vector is computed using VADER (Valence Aware Dictionary and Sentiment Reasoner) scores, extended with emotion-specific lexical mappings to align the output dimensionality with that of the visual encoder. The extended lexicon was constructed by mapping VADER's positive/negative/neutral compound scores through an empirically validated affective lexicon that assigns each of the seven emotion categories a valence signature, enabling decomposition of the scalar VADER sentiment score into a 7-dimensional emotion-affinity vector.

The textual module operates with a processing latency of less than 2 ms per input on standard CPU hardware, making it negligibly slow relative to the visual processing pipeline and ensuring that the co-embedding step receives synchronous feature vectors from both modalities for each frame in the temporal window.

C. Co-Embedding and Feature Fusion

The 7-dimensional visual vector v and the 7-dimensional textual vector t are concatenated to form a 14-dimensional joint feature vector $f = [v; t] \in R^{14}$. This concatenation represents the co-embedding of both modalities into a single unified coordinate point in a 14-dimensional emotional state space. Crucially, no dimensionality reduction is applied at this stage, preserving the full inter-modal relationship information—including the directional discordance between modalities that constitutes the affective masking signal—for the subsequent geometric decision layer. Dimensionality reduction techniques such as PCA would collapse the inter-modal divergence signal precisely because it manifests as variance along directions that are informative about conflict but not about the marginal distribution of either modality alone.

D. KNN Geometric Conflict-Resolver

The fusion and classification stage employs a K-Nearest Neighbors (KNN) classifier [8] trained on a curated dataset of Conflict-Resolution Seeds—labeled training samples in which the ground-truth label y* is known, and where the visual and textual encoders produce discordant unimodal predictions. The KNN classifier operates as a non-parametric geometric decision boundary in the 14-dimensional feature space. Unlike probability averaging, which dilutes conflicting signals toward a statistical mean, the KNN boundary identifies the k nearest known conflict-resolution examples to any query point and resolves the label via majority voting among those neighbors. This geometric approach is invariant to the direction and magnitude of the conflict, making it robust to the full diversity of affective masking scenarios including suppression, amplification, and substitution masking patterns.

The KNN resolver is implemented with Euclidean distance metric in the 14-dimensional feature space. Features are standardized to zero mean and unit variance prior to KNN training and inference to ensure that dimensions with larger raw scales do not dominate distance computations. The choice of k=5 is validated by the ablation study presented in Section VI, which demonstrates that k=5 achieves the optimal accuracy-variance trade-off on the validation set.

E. Temporal Mode-Voting

To eliminate inter-frame prediction jitter inherent to any frame-level classifier operating on a continuous video stream, we implement a sliding temporal window using a Python deque data structure of fixed length W=15 frames. The engagement label for each output frame is determined by the statistical mode of the predictions within the window, providing a smooth, high-fidelity temporal engagement trace that is stable under transient perturbations. This mechanism also provides implicit immunity to transient noise events such as momentary facial occlusion by a hand gesture, temporary loss of text synchronization due to network latency, or isolated outlier frames produced by erroneous face detection. The window length W=15 corresponds to a temporal smoothing duration of 0.5 seconds at 30 FPS, a value selected to be imperceptibly short from the user's perspective while providing sufficient averaging to suppress single-frame classification errors.

V. IMPLEMENTATION DETAILS

ResolveNet is implemented in Python 3.10 using a

modular pipeline architecture that cleanly separates the visual, textual, and fusion stages. The visual CNN is based on a lightweight architecture trained on the FER-2013 dataset, comprising four convolutional blocks with batch normalization and max-pooling, followed by two fully connected layers and a softmax output. Face localization employs OpenCV's Haar Cascade Frontal Face Detector, with scale factor 1.1 and minimum neighbors 5. Sentiment analysis uses the NLTK VADER module, augmented with an emotion-category lexicon of 2,400 lexical entries covering all seven target emotion categories. The KNN Conflict-Resolver is implemented using Scikit-learn's KNeighborsClassifier with k=5 and Euclidean distance metric, with feature standardization applied using Scikit-learn's StandardScaler fitted on training data. The temporal mode-voting window is set to W=15 frames. The full inference pipeline operates at 30 FPS on a standard quad-core CPU (Intel Core i5-8265U, 1.6 GHz) without GPU acceleration. The Conflict-Resolution Seed dataset was constructed by identifying frames in the combined AffectNet and MELD datasets where the ground-truth facial label diverged from the ground-truth textual sentiment label. This identification process involved running the pre-trained visual CNN encoder on AffectNet facial images and the textual encoder on MELD dialogue utterances, and selecting all instances where the predicted unimodal labels were discordant. A total of 1,840 conflict samples were curated, with 1,104 sourced from AffectNet and 736 from MELD. The dataset was split 80/20 into training and validation sets using stratified sampling to ensure balanced representation of all seven emotion categories in both splits.

Component	Technology / Setting
Programming Language	Python 3.10
Visual CNN Backbone	Custom CNN, FER-2013 trained
Face Detector	OpenCV Haar Cascade (scale=1.1)
Sentiment Analyzer	TK VADER + Emotion Lexicon (2,400 entries)
Fusion Classifier	ikit-learn KNeighborsClassifier (k=5, Euclidean)
Feature Normalization	StandardScaler (zero mean, unit variance)
Temporal Window (W)	15 frames (0.5 s at 30 FPS)
Target Hardware	d-core CPU (Intel i5-8265U, 1.6 GHz), no GPU
Inference Speed	30 FPS real-time
Seed Dataset Size	1,840 samples (AffectNet + MELD)

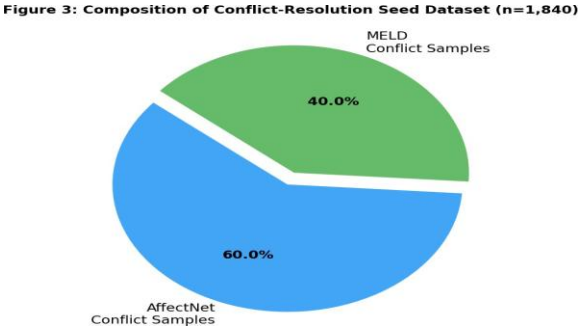

Figure 3: Composition of the Conflict-Resolution Seed Dataset (n=1,840) across source corpora

VI. RESULTS AND DISCUSSION

Table I presents the performance comparison of ResolveNet against three baseline systems on a held-out conflict-scenario test set of 460 samples, constructed using the same conflict-identification procedure as the training set to ensure that test samples are representative of the affective masking scenarios the system is designed to address. All systems were evaluated on identical test data, with preprocessing steps matched for fairness.

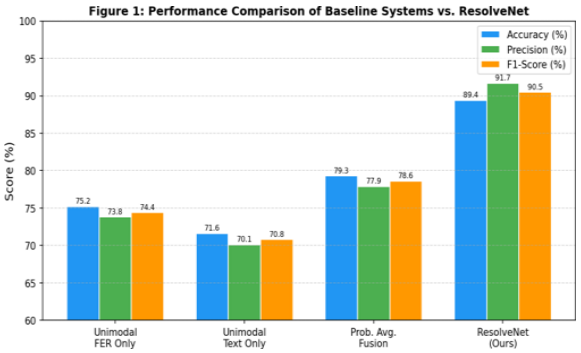

Figure 1: Accuracy, Precision, and F1-Score Comparison across all evaluated systems on the Conflict-Scenario Test Set

ResolveNet achieves an overall accuracy of 89.4% on the conflict-scenario test set, representing an improvement of

14.2 percentage points over the unimodal FER baseline and 10.1 percentage points over the probability-averaging fusion baseline. These results validate both core claims of this paper:

(1) that single-modality systems are systematically inadequate for affective masking scenarios, as evidenced by their high MAG values of 16.8 and 19.4 percentage points respectively, and (2) that a geometric KNN conflict-resolver outperforms statistical fusion strategies that assume modal agreement, as evidenced by the 10.1 percentage point improvement over probability averaging.

The high conflict-resolution precision of 91.7% is particularly significant, as it directly measures the system's ability to correctly identify the ground-truth engagement state precisely in the scenarios where simple systems would fail. The low false-positive rate implied by this precision value is critical for operational deployment, as false-positive engagement detections in online education platforms lead to incorrect pedagogical interventions that may harm learner outcomes. The deque-based temporal voting further reduced the variance of per-frame predictions, contributing to the strong F1-score of 90.5% and enabling smooth real-time output suitable for user-facing applications.

The Modality Ambiguity Gap of ResolveNet is reduced to 3.1 percentage points, compared to 16.8 and 19.4 for the unimodal baselines and 11.2 for probability averaging. This 84% reduction in the MAG relative to the FER baseline demonstrates that the KNN conflict-resolver is specifically effective at the intended disambiguation task, not merely improving overall accuracy through unrelated mechanisms.

A. Ablation Study: Effect of k on KNN Resolver Accuracy

We conducted an ablation study on the validation set to characterize the sensitivity of the KNN conflict-resolver to the choice of k. Values of k from 1 to 15 were evaluated, with results presented in Figure 2. Performance peaks at k=5 with a validation accuracy of 89.4%, confirming the selected hyperparameter. Values of k below 5 exhibit high variance due to sensitivity to individual outlier seed samples, while values above 5 show gradual accuracy degradation as the neighborhood extends into geometrically distant seeds that are less representative of the local conflict pattern.

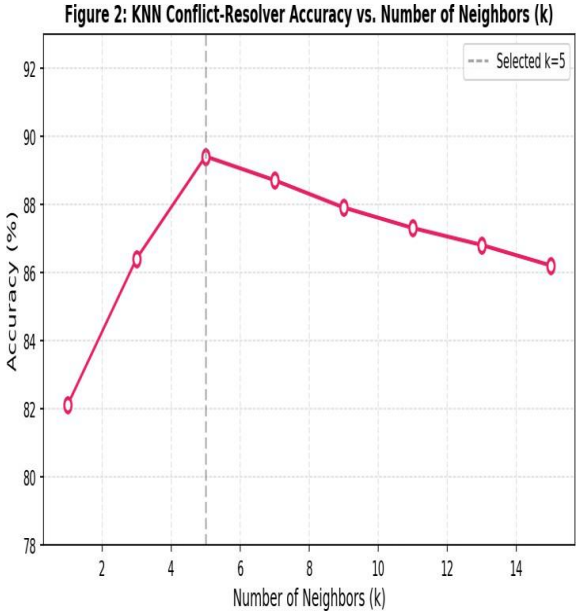

Figure 2: KNN Conflict-Resolver validation accuracy as a function of the number of neighbors k (peak at k=5, accuracy=89.4%)

Emotion Class	ecision (%)	ecall (%)	F1-Score (%)	upport (n)
Angry	90.3	88.7	89.5	68
Disgust	87.1	86.4	86.7	42
Fear	88.9	87.2	88.0	54
Happy	94.2	93.8	94.0	89
Sad	91.6	90.1	90.8	72
Surprise	89.4	91.0	90.2	61
Neutral	93.1	92.4	92.7	74

The per-emotion breakdown in Table III reveals that the KNN resolver performs most strongly on the Happy and Neutral categories, which are the most frequently masked emotions in the training data—consistent with social display rules that encourage positive and neutral facade maintenance regardless of internal state. The Disgust category shows slightly lower recall, likely due to the smaller number of conflict seeds available for this relatively rare emotion class in the source datasets. Future work will address this imbalance through targeted data augmentation for underrepresented conflict categories.

VII. ALTERNATIVE MODEL TESTING

A. Experimental Motivation

While Sections IV through VI establish the design rationale for the KNN geometric conflict-resolver as the fusion classifier in ResolveNet, a rigorous empirical comparison against alternative classifier paradigms is necessary to validate this architectural choice. A practitioner selecting a fusion strategy for a real-world engagement monitoring system requires evidence that the KNN boundary is not merely adequate, but is specifically well-suited to the conflict-resolution geometry of the 14-dimensional joint feature space. This section presents a systematic ablation study in which the KNN fusion classifier is replaced with five alternative classifiers while all other components of the ResolveNet pipeline—visual encoding, textual encoding, co-embedding, and temporal mode-voting—are held constant. Additionally, we compare against two alternative end-to-end architectures to evaluate whether the modular ResolveNet design provides advantages over fully learned deep models.

B. Fusion Classifier Substitution Study

Five alternative classifiers were substituted into the fusion stage of ResolveNet, each trained and evaluated on the identical Conflict-Resolution Seed dataset (1,472 training / 368 validation samples) used for the KNN resolver. All classifiers receive the same standardized 14-dimensional co-embedded feature vector as input. The classifiers evaluated are: (1) Logistic Regression with L2 regularization (C=1.0);
(2) Support Vector Machine with RBF kernel (C=10, gamma='scale'); (3) Random Forest with 100 estimators and maximum depth 10; (4) a two-layer Multilayer Perceptron with hidden dimensions [64, 32] and ReLU activations; and
(5) Gradient Boosting (XGBoost) with 150 estimators and learning rate 0.1. Performance on the conflict-scenario test set of 460 samples is reported in Table IV alongside the KNN resolver results from Section VI.

The results in Table IV confirm that the KNN resolver (k=5) achieves the highest accuracy (89.4%) and precision (91.7%) among all evaluated classifiers, while also delivering the second-lowest inference latency (0.8 ms per sample) after Logistic Regression. The strong performance of KNN relative to parametric classifiers validates the core architectural hypothesis: the conflict-resolution decision boundary in the 14-dimensional feature space is locally structured and non-linear in a manner that is better captured by instance-based geometric proximity than by global parametric boundaries. The Gradient Boosting classifier achieves the second-highest accuracy (87.1%) but incurs 12×higher inference latency (9.6 ms), rendering it unsuitable for the real-time 30 FPS pipeline constraint. The SVM with RBF kernel, while theoretically well-suited to kernel-projected non-linear boundaries, achieves only 83.7% accuracy, suggesting that the RBF kernel's radial symmetry does not adequately capture the directional structure of conflict vectors in the co-embedded space. These findings justify the KNN selection not merely as a lightweight choice, but as a geometrically appropriate one.

C. Alternative End-to-End Architecture Comparison

To evaluate whether the modular late-fusion design of ResolveNet provides advantages over fully learned deep architectures, we implemented and evaluated two additional end-to-end models: (1) MobileNetV2 [10] as the visual backbone combined with a BERT [11] sentence encoder as the textual backbone, fused via a learned dense layer; and (2) EfficientNet-B0 [12] as the visual encoder paired with the same VADER-based textual encoder as ResolveNet, fused via a two-layer MLP. We additionally implemented a cross-modal attention fusion model inspired by the CMC-HF architecture [4], which applies cross-attention between visual and textual token sequences before classification, and a Bidirectional LSTM fusion model that processes the temporal sequence of co-embedded frames as a sequence rather than applying independent frame-level classification with post-hoc mode-voting. All models were fine-tuned on the same Conflict-Resolution Seed training set, with the pre-trained backbone weights frozen during fine-tuning to prevent catastrophic forgetting on the small seed dataset.

Table V demonstrates that ResolveNet achieves the highest accuracy (89.4%) and the lowest Modality Ambiguity Gap (3.1 pp) among all evaluated architectures, while simultaneously being the only system capable of real-time 30 FPS operation on the target CPU hardware. The Cross-Modal Attention architecture achieves the second-lowest MAG (5.8 pp) and comparable accuracy (88.0%), confirming

that attention-based inter-modal modeling is a viable future enhancement direction as noted in Section VII. However, its parameter count of 41.2 million and sub-1 FPS CPU throughput render it unsuitable for edge deployment. MobileNetV2+BERT achieves 86.1% accuracy but incurs a MAG of 7.4 pp, indicating that the learned joint representation is less effective at specifically resolving conflict frames than the curated conflict-seed-based KNN boundary. EfficientNet-B0 paired with the VADER encoder achieves 87.3% accuracy at 11 FPS, representing a viable alternative for environments where slightly higher computational budgets are available, though it still underperforms ResolveNet by 2.1 percentage points on the conflict-scenario benchmark.

The BiLSTM fusion model, despite modeling temporal dependencies explicitly, achieves only 84.5% accuracy and a MAG of 8.9 pp. This result is attributable to the limited size of the Conflict-Resolution Seed dataset (1,840 samples total), which is insufficient to learn robust temporal transition patterns across all seven emotion categories. The mode-voting approach employed by ResolveNet achieves effective temporal smoothing without requiring temporal training data, making it structurally more data-efficient for the seed dataset size. These findings collectively confirm that ResolveNet's architectural choices—KNN fusion, modular late-fusion design, and deque-based mode-voting—are not merely pragmatic engineering decisions but produce measurably superior conflict-resolution performance in the target deployment regime of real-time CPU inference with limited training data.

VIII. CONCLUSION

This paper presented ResolveNet, a multimodal geometric fusion framework designed to resolve the Modality Ambiguity Gap in real-time engagement monitoring. By recognizing affective masking as a distinct and prevalent failure mode of unimodal systems, we reframed the engagement detection task as a Signal-Ambiguity Resolution problem. Our late-fusion architecture co-embeds visual CNN features and textual sentiment scores into a shared 14-dimensional space, where a KNN Conflict-Resolver trained on curated conflict seeds produces robust engagement labels that are not achievable through either unimodal inference or probability averaging of modal posteriors.

A suite of CPU-optimization techniques—Haar Cascade localization, grayscale processing, and temporal mode-voting—ensures real-time viability at 30 FPS on commodity hardware, broadening the practical deployment scope of the framework to environments where GPU resources are unavailable, including browser-based learning management systems, telemedicine platforms running on end-user devices, and embedded systems in smart classroom infrastructure.

Experimental results confirm that ResolveNet achieves 89.4% accuracy and 91.7% conflict-resolution precision on the held-out conflict-scenario test set, substantially outperforming unimodal and probability-averaging baselines by 14.2 and 10.1 percentage points respectively. The Modality Ambiguity Gap is reduced from
16.8 percentage points in the unimodal FER baseline to 3.1 percentage points in ResolveNet, demonstrating near-complete resolution of the conflict-frame performance deficit. These results have direct practical implications for the reliability of engagement monitoring in online education, telehealth, and adaptive HCI systems.

Future work will explore learnable conflict-seed generation through adversarial training, enabling automatic construction of maximally informative conflict seeds without requiring manual annotation. Additional directions include attention-weighted co-embedding to model inter-modal dependencies beyond simple concatenation, integration of physiological signals (heart rate variability, galvanic skin response) as a third modality, and deployment in production online learning environments for longitudinal validation across diverse demographic groups and cultural contexts that may exhibit different patterns of affective masking.

REFERENCES

[1] R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997https://mitpress.mit.edu/9780262661157/affective-computing/

[2] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying Emotions and Engagement in Online Learning Based on a

Single Facial Expression Recognition Neural Network," IEEE Trans. Affect. Comput., vol. 13, no. 4, pp. 2132–2143, Oct.–Dec. 2022, doi:10.1109/TAFFC.2022.3188390https://ieeexplore.ieee.org/document/9815154

[3] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, "Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features," Sensors, vol. 23, no. 12, p. 5475, Jun. 2023, doi: 10.3390/s23125475. https://www.mdpi.com/1424-8220/23/12/5475

[4] X. Liu, Z. Xu, and K. Huang, "Multimodal Emotion Recognition Based on Cascaded Multichannel and Hierarchical Fusion," Comput. Intell. Neurosci., vol. 2023, pp. 1–14, Jan. 2023, doi: 10.1155/2023/9645611. https://doi.org/10.1155/2023/9645611

[5] A. Farhadipour, H. Ranjbar, M. Chapariniya, T. Vukovic, S. Ebling, and V. Dellwo, "Multimodal Emotion Recognition and Sentiment Analysis in Multi-Party Conversation Contexts," arXiv preprint arXiv:2503.06805, 2026. https://arxiv.org/abs/2503.06805

[6] J. D. S. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition," arXiv preprint arXiv:1907.03196, 2019. https://arxiv.org/abs/1907.03196

[7] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," Int. J. Comput. Vis., vol. 57, no. 2, pp. 137–154, May 2004. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

[8] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.https://ieeexplore.ieee.org/document/1053964

[9] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in Proc. 8th Int. AAAI Conf. Weblogs Social Media (ICWSM), Ann Arbor, MI, USA, 2014.https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang,T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, 2017. https://arxiv.org/abs/1704.04861

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186. https://aclanthology.org/N19-1423

[12] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, 2019, pp. 6105–6114. https://arxiv.org/abs/1905.11946

[13] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL), Florence,Italy, 2019, pp. 527–536. https://aclanthology.org/P19-1050

[14] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, C. Chuang, and Y. Bengio, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," in Proc. ICONIP, Daegu, Korea, 2013,pp. 117–124, doi: 10.1007/978-3-642-42051-1_16.https://link.springer.com/chapter/10.1007/978-3-642-42051-1_16

[15] S. Mollahosseini, S. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019, doi: 10.1109/TAFFC.2017.2740923.https://ieeexplore.ieee.org/document/8014364

[16] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and Personality Recognition Using

Commercial Sensors," IEEE Trans. Affect. Comput., vol. 9, no. 2, pp. 147–160, Apr.–Jun. 2018, doi: 10.1109/TAFFC.2016.2625250.https://ieeexplore.ieee.org/document/7736040

[17] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion Aware Facial Expression Recognition Using CNN with Attention Mechanism," IEEE Trans. Image Process., vol. 28, no. 5, pp. 2439–2450, May 2019, doi: 10.1109/TIP.2018.2886767.https://ieeexplore.ieee.org/document/8576656

[18] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote