Occlusion-Robust Multimodal Emotion Recognition in VR via Fusion of Facial Images and EMG

Birgit Nierula[1,*,†], Karam Tomotaki-Dawoud[1,†], Mert Akguel[1,2], Mustafa Tevfik Lafci[1], David Przewozny[1], Anna Hilsmann[1], Peter Eisert[1,2] and Sebastian Bosse[1]

[1]*Fraunhofer HHI, Berlin, Germany*
[2]*Humboldt University of Berlin, Berlin, Germany*

Abstract

Head-mounted displays (HMDs) fundamentally limit emotion recognition in virtual reality (VR): by occluding the upper face, they render conventional image-based facial expression analysis incomplete, particularly for applications requiring real-time affective assessment. We address this challenge by fusing lower-face video with facial electromyography (EMG) from the occluded upper face to classify seven emotional categories (six basic emotions plus neutral). We introduce a synchronized multimodal dataset from 20 participants, pairing lower-face video with seven-channel upper-face EMG elicited by validated emotion stimuli. Under subject-independent test, our proposed late-fusion architecture merging convolutional visual embeddings with RBF-kernel EMG representations achieves 51% macro-F1, outperforming both image-only (41%) and EMG-only (43%) baselines. These results demonstrate that upper-face EMG provides robust complementary information under HMD-induced visual occlusion and establish a foundation for multimodal emotion recognition in naturalistic VR environments. This approach facilitates affect-adaptive applications, including communication training and therapeutic interventions. The dataset will be shared upon request under an ethical-use agreement.

Keywords

Virtual Reality, Affective Computing, Facial EMG, Multimodal Fusion, Emotion Recognition, HMD Occlusion,

1. INTRODUCTION

Virtual Reality (VR) systems have evolved beyond simple visual immersion to become platforms for human interaction and training. As these systems increasingly simulate complex social scenarios, they require mechanisms to understand and respond to users' emotional states in real-time. The immersive nature of VR enables the subjective experience of presence, the sensation of "being there" [1], which elicits authentic emotional responses comparable to real-world experiences [2, 3]. This ability to induce genuine emotional responses while maintaining full control over the scenario [4] makes VR a valuable tool for professional training applications, particularly in domains that require good communication skills, such as in healthcare [5] or law enforcement [6]. In such scenarios, automatic recognition of trainees' emotional expressions can provide objective feedback and enable adaptive role-play, supporting skill acquisition in emotionally demanding contexts.

Interpersonal communication has verbal and non-verbal aspects and one of the most important non-verbal ones is facial expressions, because they serve as one of the primary channels to convey emotional states [7] and are therefore relevant cues in a communication process. People are often unaware of their facial expressions, making assessment tools valuable for providing feedback and increasing self-awareness during communication.

Current approaches to emotion assessment in VR predominantly focus on psychophysiological measurements that capture cognitive or autonomic responses, such as changes in alpha frequency of the electroencephalogram [8], heart rate variability [9, 10, 11], or electrodermal activity [12, 13, 11]. These physiological approaches have shown success across diverse VR applications, from phobia treatment

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
*Corresponding author.
†These authors contributed equally.
Birgit.Nierula@hhi.fraunhofer.de (B. Nierula)
0000-0003-2174-5872 (B. Nierula); 0009-0006-2745-8312 (K. Tomotaki-Dawoud)

protocols [14] to therapeutic interventions [15] and presence evaluation [16]. While these physiological signals provide valuable insights into arousal and valence of the user's experienced emotional state, they do not provide information about its expression as non-verbal communication cues.

Implementing facial emotion recognition (FER) in VR poses the challenge that the head-mounted displays (HMDs) systematically occlude the upper facial region, including eyes, eyebrows and forehead that provide essential discriminative features for emotion classification. This occlusion fundamentally constrains the applicability of established camera-based facial emotion recognition methods [17], which traditionally depend on complete facial analysis to differentiate between subtle emotional distinctions, such as distinguishing surprise from fear, or anger from disgust.

The Facial Action Coding System (FACS) is a comprehensive, anatomically based framework developed by Paul Ekman and Wallace Friesen for systematically coding all observable facial movements [18]. It provides an objective, anatomy-grounded taxonomy for categorizing visible facial movements into discrete Action Units (AUs). While FACS traditionally relies on visual observation, facial electromyography (EMG) offers a complementary approach by directly measuring the underlying muscle activity of the AUs.

In this study, we propose a novel multimodal approach that combines computer vision analysis of the visible lower face with facial EMG signals captured from the occluded upper regions. We systematically evaluate this fusion strategy for classifying seven distinct emotional states, the six basic emotions introduced by Ekman [19] plus a neutral baseline. Our contributions are threefold: (i) the first synchronized dataset combining external lower-face video with upper-face EMG under VR HMD occlusion; (ii) a systematic evaluation of multimodal late-fusion strategies including kernelized vs. direct EMG encodings; and (iii) demonstration of how multimodal fusion substantially outperforms unimodal baselines in this challenging setting.

2. Related Work

Immersive virtual-reality (VR) experiences elicit genuine affect while preserving experimental control, thereby enabling ecologically valid emotion studies [12, 16]. Systematic reviews show a sharp rise in emotion-recognition studies that couple VR with physiological measures such as Electroencephalography (EEG), heart-rate variability or electro-dermal activity [12].

The challenge of FER in VR contexts has inspired several innovative technical solutions to address the fundamental problem of upper-face occlusion: (i) Optical work-arounds repurpose inward-facing cameras, for example, peri-ocular IR images from eye-tracking cameras alone are sufficient to classify a subset of five emotive expressions [20]. (ii) A complementary optical approach places photo-reflective IR sensors in the visor to capture subtle skin deformations; [21] recognized basic expressions and demonstrated real-time avatar animation. (iii) A third line reconstructs the occluded region by enabling gaze-aware, photo-realistic facial re-enactment for HMD users [22], and generative RGB-D completion removes the headset and restores identity-preserving face geometry [23]. Recent work further drives avatars from egocentric, headset-internal IR views in real time [24].

Despite growing interest in VR-based emotion recognition, the availability of comprehensive multimodal datasets remains limited. Several notable contributions have begun addressing this gap. AVDOS-VR is a database of subjective and physiological measures to affective videos recorded during a remote VR-experiment. It contains continuous valence and arousal ratings, heart response and facial EMG [25]. The emotiHeroVR Database contributes 3,556 labeled lower-face images together with 63 headset-internal expression activations from a Meta Quest Pro [26] and was collected with the VR-game EmojiHeroVR. The game gives direct feedback in form of a rewarding tone whether a facial expression indicated by an emoji was detected or not. Prior to the game, participants were trained on how to optimally perform each emotional facial expression. EmojiHeroVR thus does not extract natural emotional facial expressions but instead facial expressions that the user has learned to be detected best by the system. VREED is a multimodal affective dataset of emotional states triggered via immersive 360° video-based virtual environments in an HMD. It consists of eye tracking, electrocardiography,

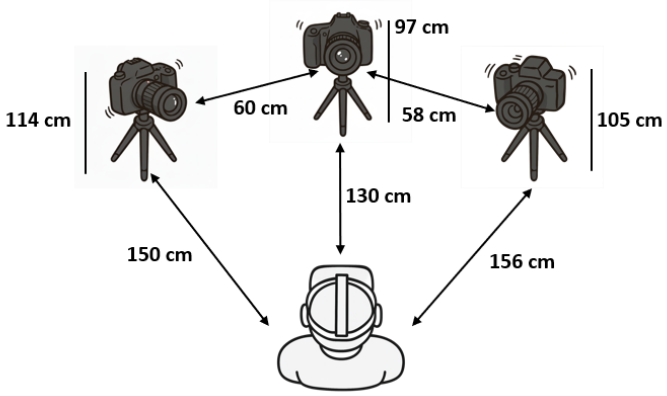

Figure 1: Experimental setup with three synchronized industrial monochrome cameras (Sony IMX547 CMOS sensor, 2448×2048 px, 22 fps) positioned at $0°$ and $\pm 30°$ relative to the sagittal plane. The monochrome design was selected for its higher sensitivity and stable performance under the low-light conditions of the immersive VR environment.

and galvanic skin response, and subjective experience reports [27]. The HEADSET collection takes a different approach, capturing emotional expressions in a volumetric camera studio with selective HMD occlusion to systematically study emotion recognition under partial visibility constraints [28].

Facial surface EMG offers an elegant solution to occlusion by directly sensing muscle activity. The emteqPRO platform implements this approach by integrating 7 facial EMG channels with photoplethysmography and inertial measurement units within a VR-compatible form factor [29, 30]. Using this mask, [31] achieved F1-macro scores up to 0.86 for five posed expressions under controlled conditions. However, their approach classified isolated facial movements rather than emotions: participants produced single-muscle actions (smile, frown, eyebrow raise, eye squeeze, and neutral), each corresponding to one primary muscle (zygomaticus major, corrugator supercilii, frontalis, orbicularis oculi). In contrast, emotions typically involve more than one muscle, even in this limited sensor setup (see for example that anger and happiness activate 2 sensors in table 1), making emotion recognition substantially more complex than detecting isolated facial movements. The same sensing technology has been successfully applied to continuous valence-arousal monitoring during affective VR video viewing [32].

The fusion of complementary sensor modalities helps overcoming individual sensor limitations. Early fusion schemes mixing forehead biopotentials with peripheral physiology improved robustness [33]. In VR, multimodal pipelines that integrate cardiovascular, electro-dermal and inertial features outperform unimodal baselines and enable explainability [34]. Deep fusion for VR emotion recognition has likewise shown benefits over single streams [35]. On the vision side, the EmojiHeroVR system reported gains when merging lower-face images with headset-internal expression activations [26].

In terms of application benefits, real-time inference can enable affect-adaptive interaction. A classic example is adapting the difficulty of *Tetris* using EEG and peripheral signals to regulate boredom, engagement, and anxiety [36].

Summary and novelty relative to prior EMG-in-VR work. Prior facial-EMG systems for VR demonstrate that upper-face EMG remains informative under HMD occlusion, but existing benchmarks often target (i) isolated facial muscle acitivity rather than multi-muscle emotion patterns, and/or (ii) continuous valence–arousal monitoring rather than discrete multi-class emotion recognition. In parallel, VR FER datasets and systems do not provide synchronized external video paired with upper-face EMG. In contrast, our work contributes (1) a synchronized dataset pairing external multi-view lower-face video with upper-face EMG under realistic HMD occlusion, (2) a systematic late-fusion evaluation for seven discrete emotion classes, and (3) an explicit comparison of direct EMG features versus kernelized EMG representations for multimodal fusion in a cross-subject setting.

Figure 2: Example frames from each monochrome camera view (left, center, right): raw frames (top) and YOLO Pose lower-face crops (bottom) at the same timestamp.

Table 1
Facial EMG channels, corresponding FACS AUs, and typical emotion associations.

Muscle (channel)	Primary AU(s)	Often seen in
Corrugator supercilii	AU4 (Brow Lowerer)	Anger, Sadness, Fear
Frontalis* (medialis)	AU1 (Inner Brow Raiser)	Sadness, Fear, Surprise
Frontalis* (lateralis)	AU2 (Outer Brow Raiser)	Surprise
Orbicularis oculi	AU6 (Cheek Raiser), AU7 (Lid Tightener)	Happiness, Anger
Zygomaticus major	AU12 (Lip Corner Puller)	Happiness

* Note the EMG sensor cannot distinguish medial and lateral frontalis.

3. Dataset

We introduce and will contribute a new synchronized multimodal VR dataset pairing lower-face video with seven-channel upper-face EMG under HMD occlusion (details on access in Sec. 7).

Participants. Twenty volunteers (10 female, mean age = 27.4 years, SD = 6.7 years) were recruited. All participants reported normal or corrected-to-normal vision.

Apparatus. Stimuli were presented via a Pico Neo 3 HMD (per-eye resolution: 2296×2408 pixels, 72 Hz refresh rate, $\sim 98^\circ$ FoV). The HMD was equipped with an EmteqPRO open-face mask [30], which synchronously recorded seven-channel facial EMG from the following muscle groups: bilateral frontalis, orbicularis oculi, zygomaticus major, and central corrugator supercilii. These muscles correspond to key FACS AUs (Table 1), providing physiological signals that complement vision-based AU features.

Lower-face video was recorded with three synchronized industrial monochrome cameras (resolution 2448×2048 px, 22 fps) based on the Sony IMX547 CMOS sensor, placed at 0° and $\pm30^\circ$ relative to the sagittal plane (Fig. 1). The monochrome design was chosen for its higher sensitivity and stable performance under the low-light conditions of our immersive VR setting. This setup captured lower-face images from three complementary viewpoints (Fig. 2, top row).

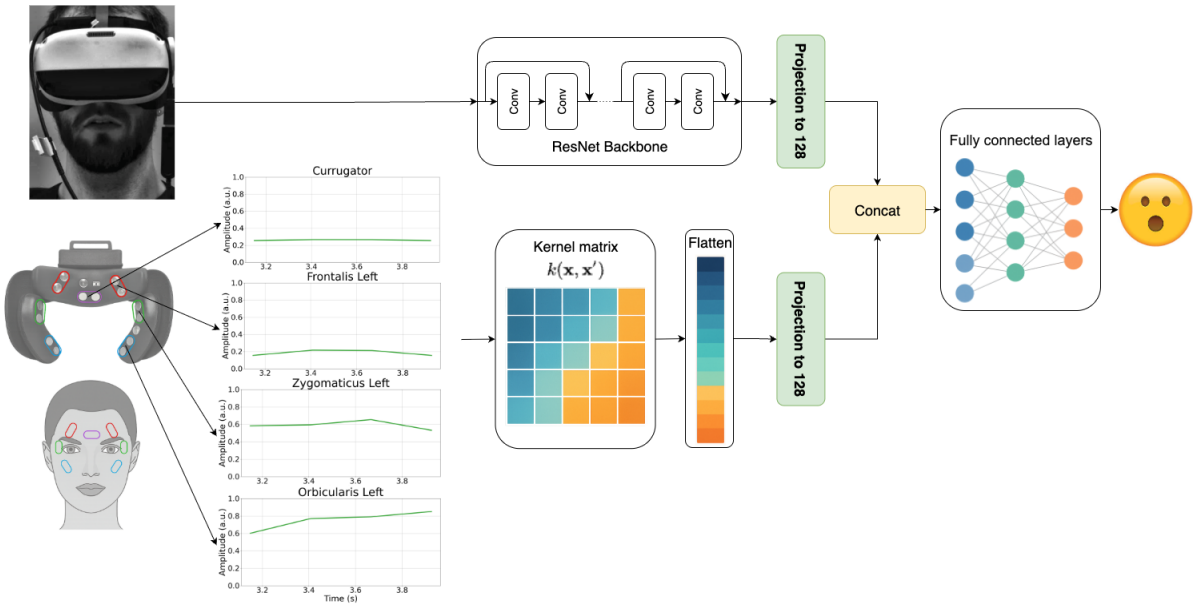

Figure 3: Overview of the proposed multimodal architecture for affect recognition in VR. The image branch (top) processes lower-face video frames using a ResNet backbone, projecting features into a 128-D embedding. The EMG branch (bottom) encodes signals from seven facial muscles into a RBF kernel matrix, which is then flattened and projected into a 128-D space. Both embeddings are concatenated and classified through fully connected layers into one of seven emotion categories. The modular design allows for the exchange of individual components: image backbones can be replaced by more modern encoding architectures (e.g., MobileNetv4 [40], Vision Transformer [41], Vision Mamba [42]); the EMG branch can utilize handcrafted features or direct post-processed signals; and additional sensor modalities can be incorporated. Similarly, the classifier head can be adapted to experiment with different fusion strategies or architectures.

Stimuli. We used 35 facial expression images from the NimStim database [37], a validated stimulus set for emotion research. Five exemplars were selected from each of seven categories (happy, sad, angry, fearful, surprised, disgusted, neutral) using intensity scores of 100% [38]. Stimuli were gender balanced (17 male, 18 female identities) and presented in randomized order. These trials were embedded within a broader experimental session that included additional tasks (not reported here).

Procedure. Following HMD and EMG sensor placement, participants performed a facial expression re-enactment task. To elicit naturalistic expressions, we employed an emotion-focused instruction set: participants were asked to first internalize the displayed emotion before expressing it facially. Each trial followed a fixed structure: (1) emotion label presentation (2 s), during which participants prepared to experience the upcoming emotion; (2) facial expression image presentation (6 s), during which participants expressed and maintained the emotion on the image on their own face; and (3) rest period displaying a fixation cross (20 s ± 4 s jitter), during which participants returned to a relaxed, neutral state. Participants were instructed to begin internalizing the emotion during the label presentation phase so that they could express it more naturally during the image presentation phase.

Synchronization. All EMG and video streams were synchronized using Lab Streaming Layer (LSL) (version:v1.16.4) [39]. The three GigE cameras were triggered by a common host, and VR event markers were inserted into the stream at trial onset.

4. Method: Multimodal Two-Stream Classification

We propose a late-fusion architecture that integrates (i) a lower-face *video stream* and (ii) an upper-face *EMG stream* to classify seven emotions under HMD occlusion (Fig. 3). The design emphasizes

modularity—either stream can be swapped or upgraded without changing the fusion contract.

4.1. EMG Stream: Sensing and Preprocessing

Signals from seven facial muscles (Tab. 1) are processed using identical parameters across participants:

1. **Filtering:** zero-phase IIR notch filter at 50 Hz and its harmonics (up to 400 Hz), followed by a 100–400 Hz band-pass filter.
2. **Epoching:** $[-1, 6]$ s relative to stimulus onset, with baseline subtraction (averaged signal amplitude from -1 to 0 s).
3. **Amplitude clipping:** global clip per muscle at the maximum observed across participant-averaged trials to limit transient spikes.
4. **Rectification:** full-wave rectification and 25 ms Root Mean Square (RMS).
5. **Temporal binning:** non-overlapping 250 ms segments, following the approach of [43].
6. **Normalization:** Min–max scaling to $[0, 1]$ performed across all trials collectively (i.e., using the global minimum and maximum across all subjects), rather than separately for each subject.

4.2. Feature Extraction (EMG)

Analysis used $[1, 6]$ s post-stimulus. Each trial was segmented with 1.0 s windows and 0.25 s hop. We used an RBF kernel representation as a compact, interpretable approximation of inter-muscle correlation structure, offering stability with limited data and low computational cost compared to deep sequence encoders. For each window, we concatenated the per-muscle scaled RMS values and computed the kernel matrix:

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2l^2}\right) \tag{1}$$

The resulting matrix is a positive semi-definite (PSD) symmetric matrix that captures non-linear correlations between the seven EMG sensors within a 1 s window. This matrix is flattened into a 784-D vector representation.

For comparison, we also evaluated direct post-processed EMG features (28-D), where each dimension corresponds to one RMS value per channel and time bin, without kernelization.

4.3. Video Stream: Lower-Face Features

For each EMG window, we take the temporally aligned final frame from each of the three cameras (Sec. 3).This pairing of three lower-face images with a single EMG window resulted in 35649 multimodal samples across the dataset. Following frame extraction, facial landmarks were localized using YOLO-Pose [44, 45] localizes landmarks; we center a 480×640 crop at the nose (Fig. 2, bottom). Frames are normalized to $[0, 1]$ and passed to a ResNet backbone initialized on the Extended Cohn-Kanade Dataset CK+ [46], providing a domain-relevant prior for facial expression recognition. Subsequently, channel-wise normalization is applied using dataset-specific statistics, with mean $\mu = [0.2948]$ and standard deviation $\sigma = [0.3002]$.

To further adapt the pre-trained model to our occlusion setting, we augmented the CK+ training images by modifying the upper-face region. Each original image was supplemented with three augmented variants (Fig. 4): (i) black occlusion patches covering the eyes, (ii) random noise patches applied to the same region, and (iii) synthetic overlays of a VR headset. These augmentations encouraged the model to focus on lower-face cues while remaining robust to varying types of occlusion.

4.4. Fusion Head and Embedding Spaces

The image branch (ResNet-18/50 truncated before its classifier) and the EMG branch (either 784-D K-EMG or 28-D direct EMG) are projected into 128-dimensional embedding spaces. We adopt a 128-dimensional projection head following the image encoder (as well as the EMG encoder), consistent with prior

(a) Original image (b) Black patch (c) Random noise patch (d) VR headset overlay

Figure 4: Examples of CK+ occlusion augmentation.

contrastive learning approaches [47, 48]. Given the size of the training set, this dimensionality provides a suitable trade-off between representational capacity, computational efficiency, and regularization; larger projection dimensions were therefore avoided to reduce the risk of overfitting. The resulting embeddings are concatenated and fed to a fully connected fusion head consisting of two hidden layers with 256 and 128 units, respectively, interleaved with Batch Normalization and ReLU activations, and a final linear layer producing seven emotion logits(Fig. 3). This late-fusion design enables independent improvements to each modality or the fusion head (e.g., alternative vision encoders, handcrafted EMG features, or different fusion operators).

4.5. Training and Evaluation

Models were trained using cross-entropy loss with Adam (lr $= 1\mathrm{e}{-3}$, batch size 8) for 30 epochs. The selection of the model for testing was based on the macro-F1 validation. Given compute constraints (35k multimodal samples across 20 identities), we fixed a subject-disjoint test set of two identities and pooled the remaining 18 identities for training/validation. Validation batches were sampled at random (fixed seed 42) from the 18 training identities and used for early stopping and model selection. In preliminary experiments, 5-fold identity splits (16 train / 2 val / 2 test) did not improve test performance compared to training on the full set of 18 identities, while incurring substantially higher training time; the larger training pool exposed the model to more diverse EMG patterns, which we found beneficial in practice.

5. Results

We first conduct an *ablation-driven model selection* to identify effective design choices for each stream and the fusion head. We then report final test results (shown in Table 2) on the held-out identities using the best configuration emerging from these ablations.

Ablation-Driven Model Selection Factors We query: (i) **modality** (Img, EMG, Img+EMG); (ii) **backbone** (ResNet-18, ResNet-50); (iii) **preprocessing** (with/without CLAHE[1] [49]); (iv) **EMG representation** (K-EMG 784-D vs. direct EMG 28-D).

Baseline performance. Conventional image-only methods performed poorly under HMD occlusion. OpenFace-3.0 [50] yielded near-chance accuracy (∼16%) with F1 below 0.1, demonstrating that established SOTA FER pipelines fail when upper-face features are unavailable. Similarly, Our ResNet+prediction head models pretrained on our occlusion-augmented CK+ data—but not fine-tuned on the target VR dataset—did not transfer effectively, confirming that domain-specific adaptation is essential.

Effect of pretraining without fine-tuning. To assess the value of domain adaptation, we evaluated ResNet-18 and ResNet-50 pretrained on CK+ (with occlusion augmentation) but not finetuned on our

[1]CLAHE is a method to adaptively improves local contrast by redistributing pixel intensities within small contextual regions while limiting noise amplification, thereby highlighting facial features more effectively. We apply CLAHE for each frame.

Table 2
Testset performance comparison of models under different preprocessing and modality settings. All values represent means across the two held-out test subjects. All ResNet models were initialized with CK+ pretraining before fine-tuning on our dataset. OpenFace and SVM baselines do not rely on CK+ pretraining. **Modality codes:** Img = images; K-EMG = kernel EMG (784-D); EMG = post-processed EMG (28-D).

Model	Finetuned (ours)	CLAHE	Modality	Accuracy	F1	Precision	Recall
OpenFace 3	–	✓	Img	0.16	0.10	0.11	0.14
OpenFace 3	–	×	Img	0.15	0.08	0.10	0.13
ResNet-18	×	✓	Img	0.13	0.082	0.060	0.142
ResNet-50	×	✓	Img	0.14	0.081	0.060	0.141
ResNet-18	×	×	Img	0.14	0.083	0.060	0.143
ResNet-50	×	×	Img	0.14	0.082	0.060	0.143
ResNet-18	✓	✓	Img	0.40	0.40	0.43	0.4
ResNet-18	✓	✓	Img+K-EMG	0.41	0.42	0.51	0.41
ResNet-50	✓	✓	Img	0.40	0.41	0.46	0.4
ResNet-50†	✓	✓	Img+K-EMG	0.50	0.51	0.55	0.50
ResNet-18	✓	×	Img	0.32	0.30	0.33	0.32
ResNet-18	✓	×	Img+K-EMG	0.39	0.35	0.38	0.39
ResNet-50	✓	×	Img	0.39	0.39	0.44	0.39
ResNet-50	✓	×	Img+K-EMG	0.40	0.40	0.43	0.40
SVM	✓	–	K-EMG	0.43	0.38	0.44	0.43
ResNet-18	✓	✓	Img+EMG	0.30	0.28	0.29	0.30
ResNet-50	✓	✓	Img+EMG	0.38	0.36	0.40	0.38
ResNet-18	✓	×	Img+EMG	0.35	0.28	0.31	0.35
ResNet-50	✓	×	Img+EMG	0.34	0.29	0.36	0.34
SVM	✓	–	EMG	0.46	0.43	0.44	0.45

† Inter-subject variability for this configuration: 1st test subject achieved F1=0.66, while the 2nd test subject achieved F1=0.36, indicating substantial individual differences in cross-subject generalization.

dataset. Performance dropped to chance level (F1 $\approx$0.08), underscoring that finetuning on domain-specific VR data is essential. This also validates our augmentation strategy: without adaptation, even occlusion-aware pretraining fails to transfer effectively.

Effect of backbone and preprocessing. With finetuning, both ResNet-18 and ResNet-50 achieved comparable performance, with ResNet-50 slightly edging over using Image-only modality, Image+(Kernel EMG or post-processed EMG). CLAHE consistently improved results by up to 10 F1 points, especially benefiting subtle contrasts in the lower-face region. Without CLAHE, performance degraded, particularly for ResNet-18.

Role of EMG. Unimodal EMG models again proved highly informative. Using direct post-processed EMG features (28-D), an SVM achieved the best unimodal score (Accuracy = 46%, F1 = 0.43), outperforming both image-only baselines and kernel-based EMG SVMs. This demonstrates that compact, temporally structured features are a viable representation, sometimes surpassing higher-dimensional kernel embeddings.

Fusion benefits. Multimodal late fusion systematically improved over unimodal baselines, but the benefits varied by EMG representation. Kernel EMG yielded the strongest overall results, with ResNet-50 + K-EMG reaching the highest performance (Accuracy = 50%, F1 = 0.51). In contrast, fusion with direct 28-D post-processed EMG provided smaller gains, peaking at F1 $\approx$ 0.36. These findings suggest that the kernelized representation more effectively captures inter-muscle correlations relevant for emotion discrimination, and the deep model is beneficial at extracting these representations.

Table 3
Training and validation performance of ResNet-18 and ResNet-50 models under different modality settings. Metrics are reported after fine-tuning on our dataset, with and without CLAHE preprocessing. Results highlight that multimodal fusion (Img+EMG or Img+K-EMG) achieves substantially higher validation accuracy and F1 compared to image-only baselines, though in some cases validation scores approach ceiling levels, suggesting potential overfitting relative to held-out test results.

Model	CLAHE	Modality	Train Acc.	Train F1	Val. Acc.	Val F1
ResNet-18	✓	Img	0.775	0.774	0.730	0.717
ResNet-18	×	Img	0.770	0.770	0.753	0.743
ResNet-50	✓	Img	0.850	0.850	0.868	0.865
ResNet-50	×	Img	0.780	0.780	0.792	0.790
ResNet-18	✓	Img+K-EMG	0.908	0.908	0.950	0.950
ResNet-18	×	Img+K-EMG	0.783	0.783	0.721	0.703
ResNet-50	✓	Img+K-EMG	0.929	0.930	0.946	0.946
ResNet-50	×	Img+K-EMG	0.788	0.788	0.818	0.817
ResNet-18	✓	Img+EMG	0.756	0.756	0.887	0.888
ResNet-18	×	Img+EMG	0.776	0.775	0.895	0.896
ResNet-50	✓	Img+EMG	0.844	0.844	0.938	0.934
ResNet-50	×	Img+EMG	0.879	0.879	0.952	0.952

Inter-subject variability. Although fusion improved average performance, substantial variability emerged between the two test subjects. For the best-performing model (ResNet-50 + K-EMG with CLAHE), F1 scores ranged from 0.36 to 0.66. Without CLAHE, both subjects showed reduced but similarly variable performance (F1: 0.34–0.46). This dispersion underscores the challenge of cross-subject generalization and motivates future work with larger participant pools.

Table 4
Illustrative per-class test accuracy for using image modality with ResNet-50 vs Multimodal ResNet-50+K-EMG (Both with CK+ pretraining, CLAHE, and finetuned on our dataset), highlighting the contribution of upper-face EMG under HMD occlusion for the different classes. Only Action Units measurable by the EMG sensors (Table 1) are listed.

Emotion	Upper-face AUs	Img (ResNet-50)	ResNet 50 Img + K-EMG
Happiness	AU6, AU12	0.64	0.75
Surprise	AU1, AU2	0.65	0.84
Anger	AU4, AU7	0.18	0.23
Fear	AU1, AU4	0.48	0.49
Sadness	AU1, AU4	0.17	0.49
Disgust	AU7	0.16	0.21
Neutral	–	0.48	0.51
Macro Avg.	—	0.40	0.51

Training and validation behavior. Table 3 provides training and validation metrics across settings. We observe that:

- Image-only models converged stably but plateaued at moderate validation scores (∼0.72–0.86 F1).
- Kernel-EMG fusion achieved very high validation performance (∼0.95 F1), but test performance did not reach the same level, indicating possible overfitting.
- Direct EMG fusion showed strong validation generalization (Val. F1 up to 0.95 with ResNet-50, no CLAHE), but these gains did not consistently translate to the test set.

A notable gap emerged between validation and held-out test performance (Tab. 3 vs. Tab. 2). Kernel-based multimodal models achieved near-ceiling validation scores (Val. F1 ≈0.95), yet their test F1

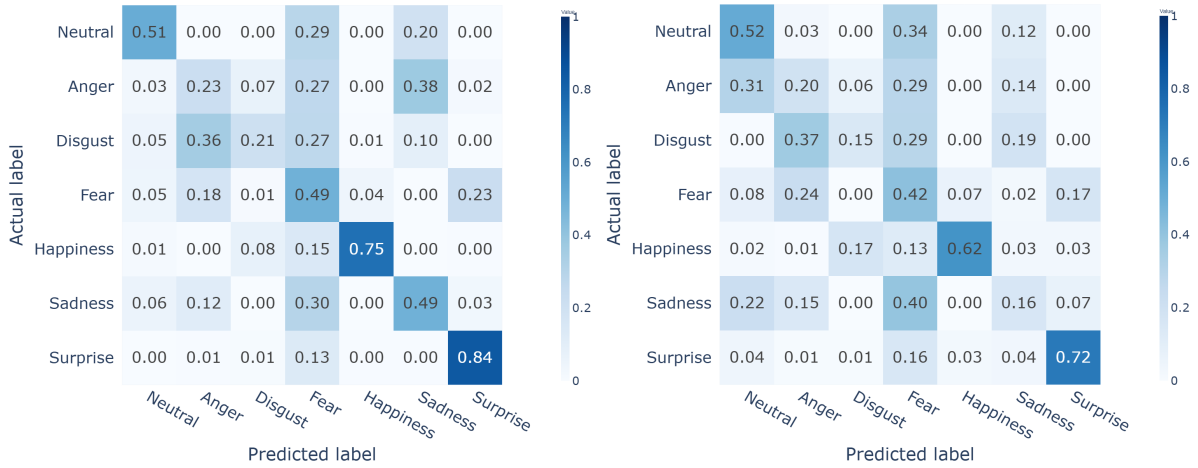

Figure 5: Normalized confusion matrices for the multimodal ResNet-50 model with CLAHE preprocessing (left) and without CLAHE (right). Both models reliably recognize highly expressive emotions such as Surprise and Happiness, but CLAHE reduces confusions between Neutral and Sadness and improves recall for subtle categories. Values are row-normalized to indicate per-class recall.

plateaued around 0.50. Similarly, fusion with direct post-processed EMG reached validation F1 above 0.93 but did not yield consistent improvements on the test set. Note that our validation split was sampled from the same 18 training identities (not disjoint by subject), which—together with the smaller validation set—likely inflates validation scores relative to the subject-disjoint test set. Thus, the observed Val–Test gap reflects both this optimistic validation protocol and the inherent difficulty of cross-subject generalization. Contributing factors include: (i) limited dataset size (35k samples across 20 participants), (ii) high intra-subject consistency that benefits validation but does not transfer to unseen identities, and (iii) the relatively low variability of lab-elicited expressions compared to spontaneous VR interactions.

These findings emphasize the need for larger-scale multimodal corpora and stronger regularization strategies (e.g., temporal modeling, dropout, or domain adaptation) to bridge the gap between in-sample validation and cross-subject generalization in immersive VR.

Per-class analysis. Table 4 reports per-class accuracy for image-only and multimodal models. Multimodal fusion with upper-face EMG yields consistent gains for emotions relying on occluded upper-face Action Units, with the largest improvements observed for Surprise (AU1/AU2: +19 points) and Sadness (AU1/AU4: +32 points). In contrast, improvements for mouth-dominant or neutral expressions are smaller, indicating that EMG primarily compensates for missing upper-face visual information under HMD occlusion rather than uniformly boosting performance across all classes.

Confusion matrix analysis. Fig. 5 presents normalized confusion matrices for the multimodal ResNet-50 with and without CLAHE, respectively. Surprise and Happiness were recognized most reliably, reaching up to 0.84 and 0.75 true positive rates with CLAHE, while Neutral was also consistently well-identified (∼0.5 recall). In contrast, Anger and Disgust were frequently confused with Fear and Sadness, suggesting overlapping muscle activation patterns in the upper face when occluded by the HMD.

Notably, the strong performance for Happiness is consistent with prior findings in human emotion perception, which show that happy expressions are detected more rapidly and robustly than other emotions due to their high visual salience, particularly driven by mouth-related features [51]. This aligns with our occlusion setting, where the lower face remains visible and informative, whereas emotions that rely more heavily on subtle upper-face cues (e.g., Anger, Disgust) remain challenging to disambiguate.

Finally, CLAHE enhanced discriminability across most classes—especially for Happiness and Surprise—while without CLAHE the model exhibited increased confusions between Sadness and Neutral.

Overall, the fusion model captures strong signals for highly salient expressions (Happiness, Surprise) but remains challenged by subtler emotions (Anger, Disgust).

Summary. Overall, our results demonstrate (i) the limited utility of image-only facial emotion recognition under HMD occlusion, even with occlusion-aware pretraining; (ii) the discriminative strength of facial EMG as a complementary modality; (iii) the superiority of kernel-based facial EMG fusion over direct 28-D features for multimodal integration; and (iv) the importance of careful preprocessing and finetuning to achieve robust performance in immersive VR settings.

6. Discussion

We demonstrate the advantages of multimodal fusion for facial emotion recognition under HMD occlusion. Conventional computer vision pipelines such as OpenFace-3, which excel on full-face datasets, collapsed to near-chance levels when applied to lower-face-only inputs in VR. This underscores the limitations of directly transferring existing facial emotion recognition methods into immersive environments.

In contrast, facial EMG signals provided a robust unimodal alternative. Despite being restricted to seven upper-face channels, EMG alone outperformed vision-only baselines and captured discriminative muscle activation patterns associated with emotional expressions. This finding is consistent with prior evidence that surface EMG remains effective under occlusion [31]. Previous electroencephalography studies in VR confirm that even low-SNR biosignals remain reliable under HMD conditions [52], further supporting the robustness of higher-SNR facial EMG in our setup. Interestingly, our ablations revealed that compact post-processed EMG (28-D RMS values) and higher-dimensional kernel embeddings (784-D) behave differently: the kernelized form provides stronger fusion benefits, while the direct representation proves more stable in unimodal settings.

The strongest contribution emerged from multimodal fusion. Combining lower-face images with EMG consistently improved recognition rates, particularly when using ResNet-50 backbones with CLAHE preprocessing. The best multimodal configuration yielded a 10-point macro-F1 gain over the corresponding image-only baseline. Confusion matrix analyses further revealed that highly expressive expressions such as Happiness and Surprise benefited the most from fusion. By contrast, subtler emotions such as Anger and Disgust remained difficult to disambiguate, often being misclassified as Fear or Sadness. These systematic confusions point to overlapping muscle activations and suggest the need for richer multimodal inputs or temporal context modeling.

From a methodological perspective, pretraining on CK+ was essential to stabilize training on our VR dataset. Without fine-tuning, pretrained networks performed at chance level, highlighting the importance of domain adaptation. Moreover, although validation metrics suggested near-ceiling performance for multimodal models (Val. F1 $\sim$0.95), generalization to unseen test identities was substantially weaker (Test F1 $\sim$0.50). This gap reflects overfitting to intra-subject consistencies and underscores the challenge of robust cross-subject affect recognition in VR.

Finally, our dataset contributes a novel resource for affective computing in immersive environments: roughly 35k paired image–EMG samples across seven emotions, collected from 20 participants under subject-exclusive splits. To our knowledge, this is the first dataset combining synchronized external video and upper-face EMG during VR exposure, and it enables systematic evaluation of multimodal facial emotion recognition under realistic occlusion.

Limitations

Evaluation protocol and generalization. We adopted a pragmatic training scheme that prioritized compute efficiency, enabling us to run a broad set of ablations within fixed resources. Rather than full LOSO or K-fold identity cross-validation, we employed a fixed subject-disjoint test set (2 identities) and trained on the remaining 18 participants. This design maximized experimental coverage while

preserving subject independence in testing. Preliminary trials with identity-fold splits did not yield higher test performance than training on the larger pool of 18 subjects, which provided the model with more diverse EMG signals in practice.

Dataset scale and demographics. Our dataset is relatively small (20 participants) and reflects a limited demographic range. Subject-specific EMG patterns are known to vary substantially, and the limited number of identities constrains the diversity of muscle activation profiles seen during training. Larger and more demographically diverse datasets are needed to better quantify robustness across users and to reduce overfitting to participant-specific idiosyncrasies. Our Dataset represents a critical first step toward tackling VR occlusion challenges. The controlled setting allowed us to tightly synchronize video and EMG, and to systematically probe multimodal fusion under HMD constraints. Importantly, the dataset design and open sharing agreement establish a foundation for future work to extend toward larger and broader demographics.

Ecological validity of emotion elicitation. Our emotion induction relied on instructed imitation of static, high-intensity NimStim facial-expression images. This design provides strong experimental control and balanced classes, but it constrains ecological validity in two important ways: (i) the captured behavior is primarily *posed* expression production rather than *spontaneous* affective display, and (ii) static stimuli and a non-interactive trial structure do not reflect the dynamics of social VR training scenarios (e.g., conversational turn-taking, speech, head motion, cognitive load, and interactive agents). Consequently, the reported performance may not directly transfer to naturalistic VR interactions without additional validation.

Limited temporal modeling. Although we apply temporal preprocessing (epoching, RMS extraction, and sliding windows), each window is classified independently and the model does not explicitly learn cross-window dynamics. This likely contributes to confusion among subtle emotions and to reduced cross-subject generalization. Incorporating lightweight temporal sequence models (e.g., LSTM/Mamba/Transformer over EMG and/or visual embeddings) and trial-level aggregation is a promising direction to improve robustness.

Finally, we observed a substantial gap between validation (Val. F1 ≈ 0.95) and subject-disjoint test performance (Test F1 ≈ 0.50). This reflects the combined effects of limited training identities, subject-specific variability in facial EMG patterns, and optimistic validation due to non-disjoint splits, rather than indicating failure of the multimodal fusion principle itself. These results highlight the inherent difficulty of cross-subject affect recognition in VR and motivate future work using larger cohorts, stricter identity-based evaluation protocols (e.g., LOSO), and explicit temporal modeling.

6.1. Conclusion

We introduced a multimodal framework for emotion recognition in VR that integrates lower-face video with upper-face EMG, directly addressing the occlusion challenge imposed by HMDs. Experiments with 20 participants demonstrated that multimodal fusion substantially outperforms unimodal baselines, with the greatest gains for salient emotions such as Happiness and Surprise.

Our contributions are threefold: (i) release of a novel VR dataset with synchronized multimodal affective recordings, (ii) analysis of kernel-based versus direct EMG encodings, and (iii) demonstration of the benefits of pretraining and multimodal fusion under HMD occlusion.

Although our focus here was methodological validation rather than deployment, the proposed architecture is computationally lightweight: ResNet backbones and compact EMG representations enable real-time inference on modern GPUs. This makes the approach directly suitable for affect-adaptive VR scenarios, such as communication training or therapy, which we aim to evaluate in future applied studies. Future work will also extend the dataset with additional physiological modalities (e.g., skin conductance, heart rate), incorporate temporal sequence modeling for more robust recognition of subtle emotions, and explore domain adaptation to reduce overfitting and improve cross-subject generalization. By bridging physiological and visual modalities, we aim to enable more robust, ecologically valid affect recognition in immersive environments.

7. ACKNOWLEDGMENTS

This research was funded by the German Ministry of Education and Research (K3VR, 13N16388), the Max Planck Society, and the Fraunhofer Society (project NeuroHum).

ETHICAL IMPACT STATEMENT

This research involves collecting and analysing biometric data, including facial expressions and EMG signals, which constitute sensitive personal information. Thus, all participants were informed about the study and its objective and provided written informed consent for participation and for the collection, analysis, and distribution of their physiological and camera data. All data were stored and analysed in pseudonymised form. The study was approved by the local ethics committee (Ethics Committee of Fraunhofer HHI, following the DFG´s Code of Conduct "Safeguarding Good Research Practice" [53]).

Beyond our study-specific safeguards, emotion recognition technologies raise significant privacy concerns as they can reveal intimate psychological states without explicit user awareness. To empower the user, consent and data management tools have been recommended [54] that proactively gather informed consent from the user and give transparent feedback when sensitive data is being recorded.

Emotion recognition systems can exhibit cultural, demographic, and individual biases that may lead to unfair or inaccurate interpretations of emotional states. Visualization dashboards with real-time insights into the data flow and the resulting decision-making may contribute to higher accuracy, trust, and unbiased performance [54].

The system lacks access to situational context, verbal content, or interaction history, all of which are essential for accurate affect interpretation. Therefore, outputs should be treated as probabilistic indicators requiring human judgment, not definitive emotional labels.

Regulatory Considerations. Emotion recognition systems based on biometric signals may fall within the scope of high-risk AI under the EU AI Act (Regulation 2024/1689), particularly when deployed in employment contexts (e.g., communication training for professionals) or healthcare settings (e.g., therapeutic interventions). Such systems may require conformity assessments, human oversight mechanisms, and transparency obligations. We emphasize that the current work constitutes a prototype intended for controlled experimental settings, not a deployment-ready product. Any translation to applied contexts would require (i) compliance with applicable AI regulations, including mandated risk assessments and conformity procedures; (ii) ethical governance through stakeholder engagement and continuous monitoring; and (iii) where research is involved, institutional review board approval for studies involving human participants.

Dataset Availability and Risk-Mitigation

We share access to the full dataset upon request; access requires acceptance of an ethical use and data-privacy-compliant agreement. The release includes synchronized lower-face video, facial EMG signals, and metadata, together with a reference split (18 training/validation identities, 2 test identities) as a starting point. Researchers are encouraged to adopt alternative evaluation protocols (e.g., LOSO, 5-fold identity CV) using the provided subject IDs and metadata; scripts for reproducing our reference split are included. To mitigate risks of misuse, the dataset is intended solely for non-commercial research on affective computing and VR interaction. Redistribution or application for surveillance, profiling, or covert monitoring is explicitly prohibited under the usage agreement.

Declaration on Generative AI

During the preparation of this work, the authors used different commercial LLMs in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] M. Slater, S. Wilbur, A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments, Presence: Teleoperators and Virtual Environments 6 (1997) 603–616. URL: https://doi.org/10.1162/pres.1997.6.6.603. doi:10.1162/pres.1997.6.6.603. arXiv:https://direct.mit.edu/pvar/article-pdf/6/6/603/1623151/pres.1997.6.6.603.pdf.

[2] A. Felnhofer, O. Kothgassner, M. Schmidt, A. Heinzle, L. Beutl, H. Hlavacs, I. Kryspin-Exner, Is virtual reality emotionally arousing? investigating five emotion inducing virtual park scenarios, International Journal of Human-Computer Studies 82 (2015) 48–56. doi:10.1016/j.ijhcs.2015.05.004, publisher Copyright: © 2015 Elsevier Ltd. All rights reserved.

[3] F. Tian, M. Hua, W. Zhang, Y. Li, X. Yang, Emotional arousal in 2d versus 3d virtual reality environments, PloS one 16 (2021) e0256211.

[4] M. V. Sanchez-Vives, M. Slater, From presence to consciousness through virtual reality, Nature reviews neuroscience 6 (2005) 332–339.

[5] M. Fernández-Alcántara, S. Escribano, R. Juliá-Sanchis, A. Castillo-López, A. Pérez-Manzano, M. Macur, S. Kalender-Smajlović, S. García-Sanjuán, M. J. Cabañero-Martínez, Virtual Simulation Tools for Communication Skills Training in Health Care Professionals: Literature Review, JMIR Medical Education 11 (2025) e63082–e63082. doi:10.2196/63082.

[6] M. Murtinger, J. C. Uhl, L. M. Atzmüller, G. Regal, M. Roither, Sound of the Police—Virtual Reality Training for Police Communication for High-Stress Operations, Multimodal Technologies and Interaction 8 (2024) 46. doi:10.3390/mti8060046.

[7] L. Sagliano, M. Ponari, M. Conson, L. Trojano, Editorial: The interpersonal effects of emotions: The influence of facial expressions on social interactions, Frontiers in Psychology 13 (2022) 1074216. doi:10.3389/fpsyg.2022.1074216.

[8] S. M. Hofmann, F. Klotzsche, A. Mariola, V. Nikulin, A. Villringer, M. Gaebler, Decoding subjective emotional arousal from eeg during an immersive virtual reality experience, elife 10 (2021) e64812.

[9] C. McCall, L. K. Hildebrandt, B. Bornemann, T. Singer, Physiophenomenology in retrospect: Memory reliably reflects physiological arousal during a prior threatening experience, Consciousness and Cognition 38 (2015) 60–70. URL: https://www.sciencedirect.com/science/article/pii/S1053810015300386. doi:10.1016/j.concog.2015.09.011.

[10] J. Marín-Morales, J. L. Higuera-Trujillo, J. Guixeres, C. Llinares, M. Alcañiz, G. Valenza, Heart rate variability analysis for the assessment of immersive emotional arousal using virtual reality: Comparing real and virtual scenarios, PloS one 16 (2021) e0254098.

[11] B. Nierula, M. T. Lafci, A. Melnik, M. Akgül, F. T. Siewe, S. Bosse, Differential physiological responses to proxemic and facial threats in virtual avatar interactions, 2025. URL: https://arxiv.org/abs/2508.10586. arXiv:2508.10586.

[12] J. Marín-Morales, C. Llinares, J. Guixeres, M. Alcañiz, Emotion recognition in immersive virtual reality: From statistics to affective computing, Sensors 20 (2020). doi:10.3390/s20185163.

[13] J. Llobera, B. Spanlang, G. Ruffini, M. Slater, Proxemics with multiple dynamic characters in an immersive virtual environment, volume 8, 2010, pp. 1–12. doi:10.1145/1857893.1857896.

[14] H. M. Peperkorn, G. W. Alpers, A. Mühlberger, Triggers of fear: Perceptual cues versus conceptual information in spider phobia, Journal of Clinical Psychology 70 (2014) 704–714. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jclp.22057. doi:10.1002/jclp.22057.

[15] J. L. Maples-Keller, B. E. Bunnell, S.-J. Kim, B. O. Rothbaum, The use of virtual reality technology in the treatment of anxiety and other psychiatric disorders, Harvard review of psychiatry 25 (2017) 103–113.

[16] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, A. Mühlberger, The impact of perception and presence on emotional reactions: a review of research in virtual reality, Frontiers in psychology 6 (2015) 26.

[17] T. Ortmann, Q. Wang, L. Putzar, Facial emotion recognition in immersive virtual reality: A systematic literature review, in: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments, 2023, pp. 77–82.

[18] P. Ekman, W. V. Friesen, Manual for the Facial Action Coding System, Consulting Psychologists Press, 1978.

[19] P. Ekman, An argument for basic emotions, Cognition and Emotion 6 (1992) 169–200. URL: https://doi.org/10.1080/02699939208411068. doi:10.1080/02699939208411068, publisher: Routledge.

[20] S. Hickson, N. Dufour, A. Sud, V. Kwatra, I. Essa, Eyemotion: Classifying facial expressions in vr using eye-tracking cameras, in: Proc. IEEE WACV, 2019, pp. 1626–1635. doi:10.1109/WACV.2019.00178.

[21] M. Murakami, K. Kikui, K. Suzuki, F. Nakamura, M. Fukuoka, K. Masai, Y. Sugiura, M. Sugimoto, Affectivehmd: facial expression recognition in head mounted display using embedded photo reflective sensors, in: ACM SIGGRAPH 2019 Emerging Technologies, SIGGRAPH '19, Association for Computing Machinery, New York, NY, USA, 2019. doi:10.1145/3305367.3335039.

[22] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, M. Nießner, Facevr: Real-time gaze-aware facial reenactment in virtual reality, ACM Trans. Graph. 37 (2018). URL: https://doi.org/10.1145/3182644. doi:10.1145/3182644.

[23] N. Numan, F. t. Haar, P. Cesar, Generative rgb-d face completion for head-mounted display removal, in: Proc. IEEE VRW, 2021, pp. 109–116. doi:10.1109/VRW52623.2021.00028.

[24] Q. Zhang, T. Xiao, H. Habeeb, L. Laich, S. Bouaziz, P. Snape, W. Zhang, M. Cioffi, P. Zhang, P. Pidlypenskyi, W. Lin, L. Ma, M. Wang, K. Li, C. Long, S. Song, M. Prazak, A. Sjoholm, A. Deogade, J. Lee, J. D. Mangas, A. Aubel, Refa: Real-time egocentric facial animations for virtual reality, in: Proc. IEEE CVPRW, 2024, pp. 4793–4802. doi:10.1109/CVPRW63382.2024.00482.

[25] M. Gnacek, L. Quintero, I. Mavridou, E. Balaguer-Ballester, T. Kostoulas, C. Nduka, E. Seiss, Avdos-vr: Affective video database with physiological signals and continuous ratings collected remotely in vr, Scientific Data 11 (2024) 132. doi:10.1038/s41597-024-02953-6.

[26] T. Ortmann, Q. Wang, L. Putzar, EmojiHeroVR: A Study on Facial Expression Recognition Under Partial Occlusion from Head-Mounted Displays , in: 2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE Computer Society, Los Alamitos, CA, USA, 2024, pp. 80–88. URL: https://doi.ieeecomputersociety.org/10.1109/ACII63134.2024.00014. doi:10.1109/ACII63134.2024.00014.

[27] L. Tabbaa, R. Searle, S. M. Bafti, M. M. Hossain, J. Intarasisrisawat, M. Glancy, C. S. Ang, Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5 (2022). URL: https://doi.org/10.1145/3495002. doi:10.1145/3495002.

[28] F. G. Lohesara, D. R. Freitas, C. Guillemot, K. Eguiazarian, S. Knorr, Headset: Human emotion awareness under partial occlusions multimodal dataset, IEEE Trans. Vis. Comput. Graph. 29 (2023) 4686–4696. doi:10.1109/TVCG.2023.3320236.

[29] H. Gjoreski, I. I. Mavridou, M. Fatoorechi, I. Kiprijanovska, M. Gjoreski, G. Cox, C. Nduka, emteqpro: Face-mounted mask for emotion recognition and affective computing, in: Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers, UbiComp/ISWC '21 Adjunct, Association for Computing Machinery, New York, NY, USA, 2021, p. 23–25. doi:10.1145/3460418.3479276.

[30] M. Gnacek, J. Broulidakis, I. Mavridou, M. Fatoorechi, E. Seiss, T. Kostoulas, E. Balaguer-Ballester, I. Kiprijanovska, C. Rosten, C. Nduka, Emteqpro—fully integrated biometric sensing array for non-invasive biomedical research in virtual reality, Frontiers in Virtual Reality 3 (2022) 781218. doi:10.3389/frvir.2022.781218.

[31] I. Kiprijanovska, B. Sazdov, M. Majstoroski, S. Stankoski, M. Gjoreski, C. Nduka, H. Gjoreski, Facial expression recognition using facial mask with emg sensors., in: VR4Health@ MUM, 2022, pp. 23–28.

[32] M. Gjoreski, I. Kiprijanovska, S. Stankoski, I. Mavridou, M. J. Broulidakis, H. Gjoreski, C. Nduka, Facial emg sensing for monitoring affect using a wearable device, Scientific Reports 12 (2022) 16876. doi:10.1038/s41598-022-21456-1.

[33] M. Khezri, M. Firoozabadi, A. R. Sharafat, Reliable emotion recognition system based on dynamic

adaptive fusion of forehead biopotentials and physiological signals, Computer Methods and Programs in Biomedicine 122 (2015) 149–164. doi:https://doi.org/10.1016/j.cmpb.2015.07.006.

[34] E. M. Polo, F. Iacomi, A. V. Rey, D. Ferraris, A. Paglialonga, R. Barbieri, Advancing emotion recognition with virtual reality: A multimodal approach using physiological signals and machine learning, Computers in Biology and Medicine 193 (2025) 110310. doi:https://doi.org/10.1016/j.compbiomed.2025.110310.

[35] P. L. Indrasiri, B. Kashyap, C. Kolambahewage, B. Nakisa, K. Ijaz, P. N. Pathirana, Vr based emotion recognition using deep multimodal fusion with biosignals across multiple anatomical domains (2024). URL: https://arxiv.org/abs/2412.02283. arXiv:2412.02283.

[36] G. Chanel, C. Rebetez, M. Bétrancourt, T. Pun, Emotion assessment from physiological signals for adaptation of game difficulty, IEEE Trans. Syst. Man Cybern. A:Syst. Hum. 41 (2011) 1052–1063. doi:10.1109/TSMCA.2011.2116000.

[37] N. Tottenham, J. Tanaka, A. Leon, T. Mccarry, M. Nurse, T. Hare, D. Marcus, A. Westerlund, B. Casey, C. Nelson, The nimstim set of facial expressions: Judgments from untrained research participants, Psychiatry research 168 (2009) 242–9. doi:10.1016/j.psychres.2008.05.006.

[38] T. F. Williams, N. Vehabovic, L. J. Simms, Developing and Validating a Facial Emotion Recognition Task With Graded Intensity, Assessment 30 (2023) 761–781. doi:10.1177/10731911211068084.

[39] C. Kothe, S. Y. Shirazi, T. Stenner, D. Medine, C. Boulay, M. I. Grivich, F. Artoni, T. Mullen, A. Delorme, S. Makeig, The lab streaming layer for synchronized multimodal recording, Imaging Neuroscience 3 (2025) IMAG.a.136. URL: https://doi.org/10.1162/IMAG.a.136. doi:10.1162/IMAG.a.136, open Access.

[40] D. Qin, C. Leichner, M. Delakis, M. Fornoni, S. Luo, F. Yang, W. Wang, C. Banbury, C. Ye, B. Akin, V. Aggarwal, T. Zhu, D. Moro, A. Howard, Mobilenetv4 – universal models for the mobile ecosystem, 2024. arXiv:2404.10518.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[42] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision mamba: efficient visual representation learning with bidirectional state space model, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, JMLR.org, 2024.

[43] J. Lou, Y. Wang, C. Nduka, M. Hamedi, I. Mavridou, F.-Y. Wang, H. Yu, Realistic facial expression reconstruction for vr hmd users, IEEE Trans. Multimed. 22 (2020) 730–743. doi:10.1109/TMM.2019.2933338.

[44] D. Maji, S. Nagori, M. Mathew, D. Poddar, Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss, 2022. URL: https://arxiv.org/abs/2204.06806. arXiv:2204.06806.

[45] G. Jocher, J. Qiu, Ultralytics yolo11, 2024. URL: https://github.com/ultralytics/ultralytics.

[46] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: Proc. IEEE CVPRW, 2010, pp. 94–101. doi:10.1109/CVPRW.2010.5543262.

[47] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 1597–1607.

[48] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[49] K. Zuiderveld, Contrast limited adaptive histogram equalization, Academic Press Professional, Inc., USA, 1994, p. 474–485.

[50] J. Hu, L. Mathur, P. P. Liang, L.-P. Morency, Openface 3.0: A lightweight multitask system for comprehensive facial behavior analysis, arXiv preprint arXiv:2506.02891 (2025).

[51] M. G. Calvo, L. Nummenmaa, Detection of emotional faces: salient physical features guide effective visual search., Journal of Experimental Psychology: General 137 (2008) 471.
[52] D. Weber, S. Hertweck, H. Alwanni, L. D. J. Fiederer, X. Wang, F. Unruh, M. Fischbach, M. E. Latoschik, T. Ball, A Structured Approach to Test the Signal Quality of Electroencephalography Measurements During Use of Head-Mounted Displays for Virtual Reality Applications, Frontiers in Neuroscience 15 (2021) 733673. doi:10.3389/fnins.2021.733673.
[53] D. Forschungsgemeinschaft, Guidelines for Safeguarding Good Research Practice. Code of Conduct (2025). doi:10.5281/ZENODO.3923601.
[54] D. Barker, M. K. R. Tippireddy, A. Farhan, B. Ahmed, Ethical Considerations in Emotion Recognition Research, Psychology International 7 (2025) 43. doi:10.3390/psycholint7020043.