D-FuseSmileNet: Enhancing spontaneous smile recognition via Hadamard fusion of handcrafted and deep-learning based features

Mohammad Junayed Hasan, Swarali Mahimkar and Prajakta Shevakari

Computer Science Department, Johns Hopkins University, Baltimore, MD, United States

ARTICLE INFO

Keywords:
Deep learning
Feature fusion
Neural networks
Spontaneous smile recognition

ABSTRACT

Distinguishing between spontaneous and posed smiles remains a challenging problem in computer vision due to their subtle spatiotemporal differences. While recent multi-task learning frameworks have demonstrated the effectiveness of combining deep learning architectures with handcrafted D-Marker features, the methods are prone to inefficiency as they require careful calibration and tuning of weights in during combination. In this paper, we propose a novel feature fusion framework, D-FuseSmileNet, that effectively integrates transformer-based representations with physiologically-grounded D-Markers through direct feature interactions, handling the limitations of the existing methods. We conduct a comprehensive evaluation of 15 feature fusion strategies, revealing that Hadamard multiplicative fusion achieves optimal performance by enabling meaningful feature interactions while preserving spatial correspondence. Our approach achieves state-of-the-art performance across multiple benchmark datasets, surpassing all previous deep-learning based methods by significant margins: UvA-NEMO: 88.7% (+0.8%), MMI: 99.7%, SPOS: 98.5% (+0.7%), BBC: 100% (+5.0%). Through computational analyses and feature visualization, we demonstrate that our fusion strategy creates more discriminative representations than existing approaches while reducing computational complexity by eliminating the need for auxiliary task supervision. The proposed method's effectiveness and computational efficiency make it particularly suitable for practical applications in human-computer interaction and affective computing.

1. Introduction

Facial expressions, as a universal form of non-verbal communication, play an essential role in the shaping of social interactions, emotional well-being, and human-computer interfaces [1, 2]. Within the spectrum of facial expressions, accurately distinguishing between spontaneous (Duchenne) and posed (non-Duchenne) smiles is a critical challenge, as it involves subtle, often imperceptible differences in muscle activations [3, 4]. The capability of reliably identifying spontaneous smiles has far-reaching implications across numerous application domains in pattern recognition and affective computing. For example, in human-computer interaction and customer service systems, responsive and emotionally intelligent agents can adapt their strategies based on the authenticity of the smile of the user, potentially improving user satisfaction and trust [5, 6]. In social robotics, spontaneous smile recognition enables robots to better gauge human emotions, thus enhancing their social presence and acceptance [7]. Beyond these domains, the recognition of spontaneous smiles can also inform psychological research, healthcare diagnostics, criminology, and marketing strategies [8, 9], underscoring its wide-ranging importance and driving a growing body of research in pattern recognition and related fields.

Early approaches for spontaneous smile recognition have often focused on hand-engineered, manually annotated features [10–19], most notably, the Duchenne Marker (D-Marker) features of the Facial Action Coding System (FACS) that highlight subtle facial muscular movements commonly associated with spontaneous or genuine emotions [20, 21]. Although effective, these hand-crafted methods are labor intensive, brittle to variations in data (e.g., illumination, pose), and often require significant domain expertise to ensure relevance and quality. With the emergence of deep learning, researchers have turned to architectures such as convolutional neural networks (CNNs), recurrent neural networks (including LSTMs), and more recently vision transformers (ViTs) [22–27], thereby achieving automatic feature extraction and surpassing traditional methods. Although these end-to-end learning strategies reduce manual efforts, they often disregard the rich, domain-relevant cues embedded in D-Markers, limiting their interpretability and potentially leaving valuable discriminative information untapped. To handle this, a multi-task learning (MTL) framework, DeepMarkerNet [28], was recently proposed. While this solution improved performance over purely deep-learning or purely handcrafted approaches, it came with several limitations. Firstly, it relies on an intricate loss-weighting scheme that must be tuned carefully, potentially complicating the training procedure. Secondly, the coupling of D-Marker prediction and smile classification may not fully exploit the complementary nature of these signals; MTL frameworks often treat D-Marker prediction as a secondary supervisory cue rather than as a core part of the feature space. Thirdly, the indirect usage of D-Markers can hinder the model's capacity to learn robust joint representations. Lastly, inference-stage complexity remains similar to other deep learning models, offering limited additional interpretability or adaptability despite the added complexity in training.

In this paper, we propose **D-FuseSmileNet**—a novel framework that directly fuses D-Marker features with deep

ORCID(s): 0009-0008-3451-0267 (M.J. Hasan)

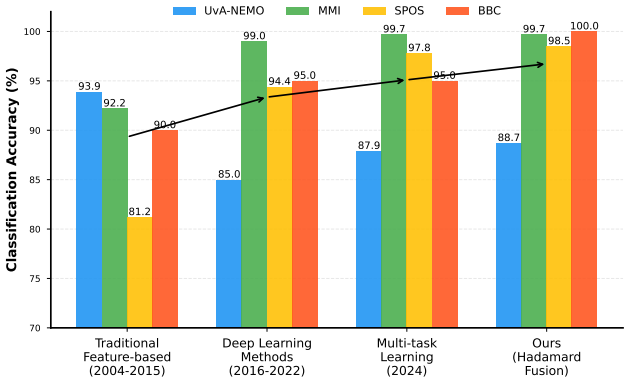

Figure 1: Gradual progression of methods for spontaneous smile recognition. Our proposed method outperforms all methods since 2016.

transformer-based representations for spontaneous smile recognition. Unlike MTL approaches that treat D-Markers as a learning signal to be predicted, we incorporate them at the feature level, allowing our model to leverage these domain-specific descriptors more explicitly. To achieve this, we systematically evaluate a comprehensive set of 15 simple and advanced feature fusion mechanisms, including various attention-based, bilinear, and multiplicative strategies. Among these, the Hadamard multiplicative fusion emerges as the most effective, offering a straightforward yet powerful means of blending domain knowledge with the richer, data-driven features learned by modern transformer architectures. By directly combining handcrafted and learned representations, our approach mitigates the training complexities, suboptimal weighting strategies, and under-realized relationships that characterize MTL frameworks. We evaluate the framework on multiple smile databases (MMI, SPOS, BBC, and UvA-NEMO), and achieve state-of-the-art performance on all of them. Fig. 1 places our proposed method in context of all exisiting methods, showing the superior performance on all deep-learning based methods since 2016.

The key contributions of this study are as follows:

- We introduce D-FuseSmileNet, a new framework that directly fuses handcrafted D-Marker features with deep transformer-based representations, moving beyond the constraints of multi-task learning.
- The proposed Hadamard multiplicative fusion outperforms all deep-learning based methods, establishing new state-of-the-art results on 4 benchmark datasets.
- We systematically investigate 15 state-of-the-art feature fusion strategies, offering insights that can benefit future research in multimodal feature fusion for facial expression analysis.

2. Related work

Existing methods for recognizing spontaneous smiles can mainly be divided into feature based methods, deep learning based methods and in addition, a combination of these two approaches.

2.1. Feature Based Methods

Feature Based Methods rely on manually designed features that are extracted from different facial regions like the eyes, lips, and cheeks. [14] used D-Markers to focus on eye muscle activation and lip corner displacement to distinguish spontaneous smiles from posed smiles. [10] analyzed the temporal dynamics of facial action units(AUs) defined in the Facial Action Coding System (FACS) to capture subtle smile variations over time. The study also identified the AUs that contained the most valuable information in spontaneous smile detection. [11] used a combination of row transformation based feature extraction algorithm that reduces computational complexity along with histogram based cascade classifier to improve performance in unconstrained environments. [13] showed how timing of smile onsets could be used to distinguish smiles. The onset amplitude of lip corner movement is less in spontaneous smiles, but the relationship between amplitude and duration is more stable. Another study [15] provides a spatiotemporal technique to distinguish between staged and spontaneous facial emotions using both natural and infrared face recordings. They extend the Completed Local Binary Patterns (CLBP) texture descriptor into the spatio-temporal CLBP-TOP features for this job by turning on temporal space and employing the image sequence as a volume.

The main limitations faced by feature based methods include the fact that a significant amount of domain expertise is required to perform manual feature engineering. These methods are also very sensitive to variations in head pose, change in lighting, and facial occlusions. They may overlook interactions between facial regions, fail to capture non-linear relationships and thus struggle with generalization across different datasets.

2.2. Deep Learning Based Methods

Deep Learning Based Methods learn the discriminative features directly from data. The use of Convolutional Neural Networks has been widespread as they can learn data without manual intervention. A study [19] used CNNs to capture high level facial representations and the Local Phase Quantization(LPQ) texture descriptor to identify subtle facial movements. The study also explored the effect of amplifying micro expressions using Eulerian Video Magnification (EVM) but it had limited impact on classification accuracy. [23] provided an end-to-end solution for the same by by training a series of convolution layers followed by a ConvLSTM layer from scratch. Another study introduced "MeshSmileNet" [24] to prevent the inclusion of irrelevant facial features, by using facial features extracted using the Attention Mesh model. The model analyzed spatial relativity by grouping nearby landmarks to capture local facial interactions and temporal trajectories using a self-attention mechanism to track landmark movements over time. The transformer architecture introduced in [27], where a Vision Transformer (ViT) could be used to divide each image in

patches of 16x16 pixels which are then used as input tokens. These can be used for classification directly and surpassed CNNs on large datasets.

Limitations Identified with the Deep Learning Based Methods like the ViT [27] require large scale datasets for pre-training to achieve competitive performance. This also requires substantial computational resources. These methods are also prone to overfiiting by learning spurious correlations and irrelevant features. They also tend to underutilize domain-specific insights like D-Markers that could improve interpretability.

2.3. Combined Methods

A novel multitask approach that combined feature based methods with deep learning based methods was introduced in [28]. A Relativity Network was used to capture spatial relationships between the facial landmarks and a Trajectory Network was used for dynamics using self-attention. A binary cross-entropy loss function was used for smile classification and mean squared error loss was used to supervise D-Marker prediction. D-markers were not computed during inference, which made the framework efficient. However, the losses were combined using a weighted combination of two losses, which required heavy experimentation and hyperparameter tuning, making the training process cumbersome.

3. Methodology

3.1. Problem formulation

Let $\{\mathbf{X}_i, y_i\}_{i=1}^{N}$ denote a collection of N video samples, where each sample $\mathbf{X}_i = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{T_i}\}$ is a sequence of T_i facial frames. The label $y_i \in \{0, 1\}$ indicates whether the i-th video contains a posed ($y_i = 0$) or a spontaneous ($y_i = 1$) smile. Our primary objective is to learn a classification function $f : \mathbf{X}_i \mapsto y_i$ that accurately discriminates between these two classes.

In addition to the raw video frames, we assume the availability of a handcrafted feature vector $\mathbf{Z}_i \in \mathbb{R}^M$ for each video i. This $\mathbf{Z}_i$ encodes D-Marker features—highly discriminative handcrafted descriptors known to capture subtle facial muscle activations associated with spontaneous smiles. While previous approaches have either relied on these D-Markers as direct inputs or as auxiliary supervisory signals in a multi-task setting, our goal is to integrate these features more directly at the representation level.

Formally, let $\mathcal{F}$ be a transformer-based feature extractor that processes the raw frames $\mathbf{X}_i$ to produce a learned visual representation $\mathbf{H}_i \in \mathbb{R}^D$:

$$\mathbf{H}_i = \mathcal{F}(\mathbf{X}_i). \tag{1}$$

We define a fusion operator $\otimes : \mathbb{R}^D \times \mathbb{R}^M \mapsto \mathbb{R}^Q$ that combines the learned representation $\mathbf{H}_i$ and the handcrafted D-Marker features $\mathbf{Z}_i$, producing a fused feature vector $\mathbf{F}_i \in \mathbb{R}^Q$:

$$\mathbf{F}_i = \mathbf{H}_i \otimes \mathbf{Z}_i. \tag{2}$$

Our classifier $C : \mathbb{R}^Q \mapsto [0, 1]$ then predicts the probability of the smile being spontaneous:

$$\hat{y}_i = C(\mathbf{F}_i). \tag{3}$$

The fusion operator $\otimes$ can be implemented using various strategies, such as concatenation, attention-based weighting, bilinear pooling, or element-wise multiplicative integration. In this work, we systematically explore multiple such fusion techniques and identify those that most effectively combine the complementary information from $\mathbf{H}_i$ and $\mathbf{Z}_i$.

Training the model involves optimizing the parameters of $\mathcal{F}$, $\otimes$, and C to minimize a suitable loss function $\mathcal{L}$, which in our case is the binary cross-entropy:

$$\mathcal{L}(\hat{y}_i, y_i) = -\big[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\big]. \tag{4}$$

By jointly learning $\mathcal{F}$, $\otimes$, and C, the model can exploit both learned representations from a powerful transformer backbone and domain-specific D-Marker cues. This direct fusion paradigm eschews the complexities and constraints of multi-task frameworks, paving the way for more interpretable, efficient, and effective spontaneous smile classification models.

3.2. Handcrafted feature extraction

The handcrafted D-Marker features are extracted using an effective study [17]. The process involves selection of facial landmarks, preprocessing and preparing them for D-Marker calculation, and aggregating features from three facial regions: eyes, cheeks and lips.

Landmark selection and preprocessing. Raw video samples $\mathbf{X}_i$ are passed through an automated facial mesh prediction model, AttentionMesh [29]. The AttentionMesh model extracts 478 3D facial landmark points through an automatic tracking process. From these landmark points, following an effective method , we isolate a small subset of $M = 11$ key points that best capture the contraction and movement of facial muscles orbicularis oculi and zygomaticus major, as well as subtle cheek and lip-corner motions. These key points include corners and midpoints around the eyes, cheeks, and lip corners, as well as a nose tip reference point for spatial alignment. Table 1 provides the mapping of the selected landmarks to their corresponding indices in the 478 facial landmark points of the AttentionMesh landmark extractor. We use indices 1 to 11 for these points respectively in the rest of the paper.

To ensure consistent geometric interpretation, the selected landmarks of each frame are aligned to a reference coordinate system. Let $\mathbf{p}_j^{(x)} \in \mathbb{R}^3$ denote the 3D coordinates of the j-th selected landmark in frame x. We compute a plane using eye and nose reference points, and derive a normal vector to this plane. Using this normal, we estimate and remove head rotations (roll, yaw, pitch), followed by scale and translation adjustments. This normalization ensures that extracted features are robust to head pose variations and camera viewpoints.

Table 1
Mapping of the facial feature points to the indices of the AttentionMesh facial landmark extractor.

Facial Point	Attention Mesh Index
Right-eye outer corner	33
Right-eye center	159
Right-eye inner corner	133
Left-eye outer corner	362
Left-eye center	386
Left-eye inner corner	263
Right cheek	50
Left cheek	280
Nose tip	1
Right lip-corner	62
Left lip-corner	308

D-Marker computation. Once the landmarks are normalized, we derive three key dynamic measurements that collectively capture the structure of a smile over time:

1. *Lip Dynamics* (D_{lip}): Measures the relative distance and angular changes of lip corners from a stable reference, reflecting mouth opening and lip pulling (Eq. 5).

2. *Eye Aperture* (D_{eye}): Quantifies eyelid opening or closing, capturing the hallmark crinkling around the eyes associated with spontaneous smiles (Eq. 6).

3. *Cheek Elevation* (D_{cheek}): Tracks the vertical displacement of cheek regions, which lift prominently during spontaneous smiles (Eq. 7).

$$D_{\text{lip}}(x) = \frac{\gamma(\frac{p^1_{10}+p^1_{11}}{2}, p^x_{10}) + \gamma(\frac{p^1_{10}+p^1_{11}}{2}, p^x_{11})}{2\gamma(p^1_{10}, p^1_{11})}, \tag{5}$$

$$D_{\text{eye}}(x) = \frac{\Gamma(\frac{p^x_1+p^x_3}{2}, p^x_2)\gamma(\frac{p^x_1+p^x_3}{2}, p^x_2) + \Gamma(\frac{p^x_4+p^x_6}{2}, p^x_5)\gamma(\frac{p^x_4+p^x_6}{2}, p^x_5)}{2\gamma(p^x_1, p^x_3)}, \tag{6}$$

$$D_{\text{cheek}}(x) = \frac{\gamma(\frac{p^1_7+p^1_8}{2}, p^x_7) + \gamma(\frac{p^1_7+p^1_8}{2}, p^x_8)}{2\gamma(p^1_7, p^1_8)}, \tag{7}$$

where, $\mathbf{p^x_i}$ represents the landmark at index i in frame x, $\gamma()$ represents the Euclidean distance, and $\Gamma(\mathbf{p_i}, \mathbf{p_j})$ represents the relative vertical location, which equals -1 if $\mathbf{p_j}$ is located vertically below $\mathbf{p_i}$ on the face, and 1 otherwise.

For each frame x, these metrics $D_{\text{lip}}(x)$, $D_{\text{eye}}(x)$, and $D_{\text{cheek}}(x)$ are computed relative to baseline configurations observed in the initial frames. Differences in these values over time encode the temporal evolution of the smile.

Temporal Feature Aggregation. The three key phases of the smile, such as longest increasing segment (onset), stable apex interval, and longest decreasing segment (offset), are identified from the D-Marker metrics using the approach proposed in [30]. Each phase is again divided into increasing and decreasing segments for detailed analyses. From these segments, we derive a comprehensive set of temporal descriptors: duration-related measures, amplitude magnitudes, velocity and acceleration cues, and various ratios that capture the pattern of smile in the videos. As shown in Table 2, a total of 25 features are calculated for each of the three phases, giving 75 features for each of the three facial regions (eyes, lips and cheeks). By concatenating all these features we form a final k-dimensional D-Marker feature vector, $\mathbf{D_i} = \{\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_k}\}$ where $k = 225$ in our implementation. This hand-crafted D-Marker representation encodes the geometric and temporal patterns of spontaneous smiles in a structured, low-dimensional feature space. Although hand-crafted, these features offer complementary information to data-driven embeddings extracted by deep architectures, paving the way for effective feature-level integration in our proposed framework.

Table 2
Facial feature definitions and total number of features extracted for each group.

Feature	Definition	Features
Duration	$\left[\frac{\eta(\mathcal{D}^+)}{\omega}, \frac{\eta(\mathcal{D}^-)}{\omega}, \frac{\eta(\mathcal{D})}{\omega}\right]$	3
Duration Ratio	$\left[\frac{\eta(\mathcal{D}^+)}{\eta(\mathcal{D})}, \frac{\eta(\mathcal{D}^-)}{\eta(\mathcal{D})}\right]$	2
Maximum Amplitude	$\max(\mathcal{D})$	1
Mean Amplitude	$\left[\frac{\sum\mathcal{D}}{\eta(\mathcal{D})}, \frac{\sum\mathcal{D}^+}{\eta(\mathcal{D}^+)}, \frac{\sum\lvert\mathcal{D}^-\rvert}{\eta(\mathcal{D}^-)}\right]$	3
STD of Amplitude	$\text{std}(\mathcal{D})$	1
Total Amplitude	$\left[\sum\mathcal{D}^+, \sum\lvert\mathcal{D}^-\rvert\right]$	2
Net Amplitude	$\sum\mathcal{D}^+ - \sum\lvert\mathcal{D}^-\rvert$	1
Amplitude Ratio	$\left[\frac{\sum\mathcal{D}^+}{\sum\mathcal{D}^+ + \sum\lvert\mathcal{D}^-\rvert}, \frac{\sum\lvert\mathcal{D}^-\rvert}{\sum\mathcal{D}^+ + \sum\lvert\mathcal{D}^-\rvert}\right]$	2
Maximum Speed	$\left[\max(\mathcal{V}^+), \max(\lvert\mathcal{V}^-\rvert)\right]$	2
Mean Speed	$\left[\frac{\sum\mathcal{V}^+}{\eta(\mathcal{V}^+)}, \frac{\sum\lvert\mathcal{V}^-\rvert}{\eta(\mathcal{V}^-)}\right]$	2
Max. Acceleration	$\left[\max(\mathcal{A}^+), \max(\lvert\mathcal{A}^-\rvert)\right]$	2
Mean Acceleration	$\left[\frac{\sum\mathcal{A}^+}{\eta(\mathcal{A}^+)}, \frac{\sum\lvert\mathcal{A}^-\rvert}{\eta(\mathcal{A}^-)}\right]$	2
Ampl. Duration Ratio	$\frac{(\sum\mathcal{D}^+ - \sum\lvert\mathcal{D}^-\rvert)\omega}{\eta(\mathcal{D})}$	1
Ampl. Difference	$\frac{\lvert\sum\mathcal{D}_L - \sum\mathcal{D}_R\rvert}{\eta(\mathcal{D})}$	1
	Total Number of Features:	**25**

3.3. D-FuseSmileNet Architecture

The architecture diagram of the proposed method is illustrated in Fig. 2. The architecture mainly has two parts: the D-Marker extraction using manual handcrafted approaches, as detailed in the previous section, and the automatic deep-learning part.

Automatic Feature Extraction. To obtain robust automatic representations of facial dynamics, we adopt the current state-of-the-art transformer based architecture, MeshSmileNet [24]. The transformer network first extracts 3D facial landmarks from each frame of the input video sequence, thereby condensing raw pixel information into geometric coordinates. These landmarks are then processed through a curve-based encoder, which groups semantically related points into curves for enhanced spatial reasoning.

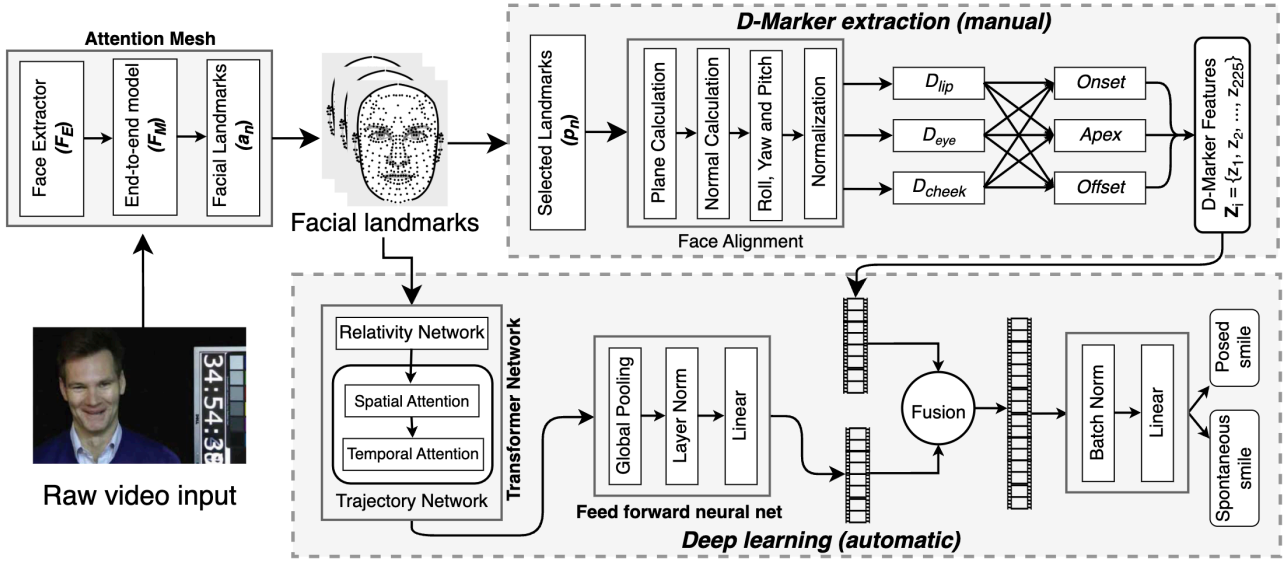

Figure 2: Overview of our proposed framework for spontaneous smile recognition. The pipeline consists of two parallel streams: (1) an automatic feature extraction path using Attention Mesh for facial landmark detection followed by transformer-based feature learning, and (2) a D-Marker extraction path that computes handcrafted physiological features. These complementary features are combined through various fusion mechanisms to achieve robust smile classification.

Subsequently, a Vision Transformer (ViT)-style architecture processes the temporal evolution of these curve embeddings. Self-attention layers capture complex spatial-temporal dependencies, and the final output is an embedding vector $\mathbf{H} \in \mathbb{R}^D$ that encodes the temporal progression of facial movements. This embedding serves as our "automatic feature" representation, containing rich, data-driven cues for distinguishing subtle spontaneous and posed smiles.

Feature Fusion with D-Markers. Having established a robust automatic embedding $\mathbf{H}$ from the transformer-based feature extractor and a handcrafted D-Marker representation $\mathbf{Z}$ as described in Section 3.2, our objective is now to integrate these complementary sources of information at the feature level. By fusing learned and handcrafted features directly—rather than treating D-Marker supervision as an auxiliary task—we aim to realize a more synergistic representation that leverages both domain knowledge and data-driven patterns. Below, we introduce a diverse set of fusion techniques, each offering a distinct strategy for blending $\mathbf{H}$ and $\mathbf{Z}$. For simplicity, let $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{z} \in \mathbb{R}^D$ be feature vectors obtained by projecting $\mathbf{H}$ and $\mathbf{Z}$ into a common dimension.

(1) Concatenation. We first project the transformer features $\mathbf{x}$ and D-Marker features $\mathbf{z}$ to a common 256-dimensional space using linear projections, followed by concatenation and non-linear dimensionality reduction:

$$\mathbf{F} = \phi([\mathbf{x}; \mathbf{z}]), \tag{8}$$

where ϕ is implemented as a four-layer MLP (256→256→256→256) with ReLU activations, layer normalization, and dropout (p=0.1) between each layer. This architecture allows for deep feature interaction while maintaining regularization.

(2) Gated Concatenation. This method introduces an adaptive gating mechanism implemented through a dedicated neural network pathway:

$$\mathbf{g} = \sigma(W_g[\mathbf{x}; \mathbf{z}]), \quad \mathbf{F} = \mathbf{g} \odot \mathbf{x} + (1 - \mathbf{g}) \odot \mathbf{z}, \tag{9}$$

where W_g is implemented as a linear layer (512→256) followed by sigmoid activation. The gated features undergo further refinement through a two-layer MLP (256→256→256) with ReLU activations and dropout, allowing the network to learn complex feature relationships post-gating.

(3) Additive Fusion. We implement a sophisticated attention mechanism with separate feature processing streams:

$$\mathbf{x}_{proj} = \text{MLP}_x(\mathbf{x}), \quad \mathbf{z}_{proj} = \text{MLP}_z(\mathbf{z}), \tag{10}$$

$$\boldsymbol{\alpha} = \text{softmax}(W_a[\mathbf{x}_{proj}; \mathbf{z}_{proj}]), \quad \mathbf{F} = \alpha_1 \mathbf{x}_{proj} + \alpha_2 \mathbf{z}_{proj}, \tag{11}$$

where each MLP consists of a linear layer, layer normalization, ReLU, and dropout. The fusion weights are computed using a dedicated linear layer (512→2) with softmax normalization, ensuring proper weighting of each feature stream.

(4) Multiplicative (Hadamard) Fusion. We implement this method with careful feature preprocessing and post-fusion refinement:

$$\mathbf{x}_{proj} = W_x \mathbf{x}, \quad \mathbf{z}_{proj} = W_z \mathbf{z}, \quad \mathbf{F} = \mathbf{x}_{proj} \odot \mathbf{z}_{proj}, \tag{12}$$

where W_x and W_z are linear projections (256→256), followed by a two-layer MLP (256→256→256) with ReLU activations and dropout for post-fusion feature refinement.

(5) Attention-based Fusion. We implement scaled dot-product attention with learned projections:

$$\mathbf{F} = \text{softmax}\left(\frac{(\mathbf{x}W_Q)(\mathbf{z}W_K)^T}{\sqrt{256}} \right) (\mathbf{z}W_V), \tag{13}$$

where W_Q, W_K, W_V are implemented as separate linear layers (256→256). To handle single token inputs, we reshape features to [batch_size, 1, 256] before attention computation and squeeze afterward.

(6) Multi-head Attention Fusion. We extend the attention mechanism to multiple heads:

$$\mathbf{F} = \text{Concat}_{h=1}^{4} \left[\text{Attention}_h(\mathbf{x}, \mathbf{z})\right] W_O, \tag{14}$$

with 4 attention heads, each operating on 64-dimensional feature subspaces (256/4). W_O is implemented as a linear layer (256→256) followed by dropout (p=0.1).

(7) Cross Attention. We implement bidirectional attention flow using separate Q/K/V projections:

$$Q = W_Q\mathbf{x}, \quad K = W_K\mathbf{z}, \quad V = W_V\mathbf{z}, \tag{15}$$

$$\mathbf{F} = \text{softmax}\left(\frac{QK^T}{\sqrt{256}}\right)V, \tag{16}$$

where all projections are linear layers (256→256) with proper reshaping to [batch_size, 1, 256] for attention computation.

(8) Multi-head Cross Attention. Extends cross-attention with parallel processing streams:

$$\text{Head}_i = \text{Attention}(W_Q^i\mathbf{x}, W_K^i\mathbf{z}, W_V^i\mathbf{z}), \tag{17}$$

$$\mathbf{F} = W_O[\text{Concat}(\text{Head}_1, ..., \text{Head}_4)], \tag{18}$$

implemented with 4 heads, each with its own Q/K/V projections (256→64) and final output projection W_O (256→256).

(9) FiLM. We implement feature-wise linear modulation through a dedicated conditioning network:

$$[\boldsymbol{\gamma}, \boldsymbol{\beta}] = \text{MLP}(\mathbf{z}), \quad \mathbf{F} = \boldsymbol{\gamma} \odot \text{LN}(\mathbf{x}) + \boldsymbol{\beta}, \tag{19}$$

where MLP is a two-layer network (256→512→512) with ReLU activation, generating both scaling and shifting parameters. LN denotes layer normalization applied to transformer features.

(10) FiLM + Hadamard. Combines FiLM conditioning with direct multiplication:

$$\mathbf{F} = (\boldsymbol{\gamma} \odot \text{LN}(\mathbf{x}) + \boldsymbol{\beta}) + (\mathbf{x} \odot \mathbf{z}), \tag{20}$$

followed by a refinement MLP (256→256→256) with layer normalization and ReLU activation.

(11) Bilinear Pooling. We implement full bilinear interactions with dimensionality reduction in the following way:

$$\mathbf{F} = W_o((\mathbf{x}W_1)(\mathbf{z}W_2)^T), \tag{21}$$

where W_1, W_2 are linear projections (256→256) and W_o reduces the bilinear feature to 256 dimensions.

(12) Bilinear + Hadamard. We combine bilinear and element-wise interactions in the following way:

$$\mathbf{F} = \text{MLP}(\text{Bilinear}(\mathbf{x}, \mathbf{z}) + (\mathbf{x} \odot \mathbf{z})), \tag{22}$$

where MLP (256→256→256) with ReLU and layer normalization refines the combined features.

(13) Factorized Bilinear Pooling. We employ an efficient low-rank approximation technique, as shown below:

$$\mathbf{F} = W_o((\mathbf{x}W_p) \odot (\mathbf{z}W_q)), \tag{23}$$

where W_p, W_q project to a 512-dimensional intermediate space and W_o projects back to 256 dimensions.

(14) Factorized Bilinear + Hadamard. This method combines factorized bilinear pooling with direct multiplication as follows:

$$\mathbf{F} = \text{MLP}(\text{FactorizedBilinear}(\mathbf{x}, \mathbf{z}) + (\mathbf{x} \odot \mathbf{z})), \tag{24}$$

using a combine layer (256→256) with ReLU activation followed by the standard fusion MLP.

(15) Gated Concatenation + Hadamard. In this hybrid approach, we combine three interaction paths:

$$\mathbf{F} = (\mathbf{x} \odot \mathbf{z}) + \left(\sigma(W_g[\mathbf{x}; \mathbf{z}]) \odot \mathbf{x} + (1 - \sigma(W_g[\mathbf{x}; \mathbf{z}])) \odot \mathbf{z}\right), \tag{25}$$

where each component uses the same architecture as its standalone counterpart, and the final features undergo refinement through the fusion MLP.

All the fusion strategies are integrated into the D-FuseSmileNet pipeline, and trained end-to-end. By exploring this broad spectrum of fusion operators, we identify effective mechanisms that fully exploit the complementary nature of automatic deep embeddings and handcrafted D-Marker features.

4. Experiments

4.1. Setup

Datasets. We evaluate our proposed method on four widely-used benchmark datasets: **(i)** The UvA-NEMO smile database [17] is the most comprehensive dataset for smile authenticity analysis, comprising 1240 high-definition videos (597 spontaneous, 643 posed) recorded at 1920×1080 resolution and 50 FPS. The dataset features 400 subjects (185 female, 215 male) across a broad age spectrum (8-76 years), with controlled illumination conditions enabling focus on expression dynamics. **(ii)** The BBC dataset presents unique challenges through its collection of celebrity interviews, containing 20 smile videos (10 spontaneous, 10 posed) recorded in real-world conditions. The lower resolution (314×286) and varying illumination make it particularly suitable for evaluating robustness. **(iii)** The MMI facial expression dataset contributes 187 smile videos (138 spontaneous at 640×480/29 FPS, 49 posed at 720×576/25 FPS), offering diversity in recording conditions. **(iv)** The SPOS dataset [15] includes grayscale sequences captured at 640×480 resolution and 25 FPS. We utilize its gray-scale smile sequences, which comprise both spontaneous and posed expressions under controlled settings. These datasets present varying challenges through their different recording conditions, resolutions, and subject demographics, enabling comprehensive evaluation of our method's generalization capabilities. Notably, while UvA-NEMO offers ideal conditions for analyzing subtle expression differences, the BBC dataset tests robustness to real-world variations, and MMI and SPOS provide additional validation across different video qualities and frame rates. The dataset details are summarized in Table 3. Some data samples randomly selected from the UvA-NEMO database can be

Figure 3: Data samples randomly drawn from the UvA-NEMO database showing neutral face (top), posed enjoyment smile (middle), and spontaneous enjoyment smile (bottom).

Table 3
Details of the benchmark datasets used in our experiments.

Database	Number of Subjects			Video		Annotation	Age Range
	Spon.	Posed	Total	Resolution	FPS		
UvA-NEMO	357	368	400	1920 × 1080	50	Yes	8-76
MMI	25	30	55	720 × 576 640 × 480	25 29	Yes	19-64
SPOS	7	7	7	640 × 480	25	No	Unknown
BBC	10	10	20	314 × 286	25	No	Unknown

visualized in Fig. 3

Evaluation Protocol. We adopt rigorous cross-validation protocols for comprehensive evaluation across datasets. For UvA-NEMO, we employ 10-fold cross-validation following established protocols [17]. The BBC, MMI, and SPOS datasets are evaluated using 10-fold, 9-fold, and 7-fold cross-validation respectively, ensuring fair comparison with previous works [16]. To ensure robust performance estimation, we conduct ten independent runs for each dataset, with each subset serving as test data exactly once. Special care is taken to maintain subject independence across train-test splits, preventing potential data leakage. Performance is measured using **classification accuracy** averaged across all folds.

Implementation Details[1]**.** Our framework is implemented in PyTorch and trained on an NVIDIA Tesla T4 GPU on the Amazon AWS EC2 server. The preprocessing pipeline extracts 478 3D facial landmarks using Attention Mesh, which are subsequently transformed into fixed-length sequences of 16 frames for consistent input dimensionality. The network architecture consists of a temporal model with a CurveNet encoder, followed by our proposed fusion module that combines transformer features with D-Marker representations. Training proceeds for 300 epochs with a batch size of 16, using the AdamW optimizer with an initial learning rate of 5e-4. The spatial-temporal transformer employs 6 blocks for spatial attention and 3 blocks for temporal modeling, each with 4 attention heads. We utilize binary cross-entropy loss for smile classification, with all network weights initialized using He initialization. Layer normalization and dropout (p=0.1) are applied throughout the network to prevent overfitting. During inference, we apply temporal average pooling over predictions to obtain the final classification result.

[1]Codes and models are available at: https://github.com/junayed-hasan/smile-recognition-fusion/

4.2. Results with fusion techniques

To comprehensively evaluate different feature fusion strategies, we conducted extensive experiments across four benchmark datasets. Table 4 presents the classification accuracy for each fusion technique.

The results reveal several crucial observations: ***(i)*** The Hadamard multiplicative fusion consistently demonstrates superior performance, achieving optimal or near-optimal results across all datasets (average accuracy: 96.73%). This suggests that element-wise multiplication effectively preserves the spatial correspondence between transformer features and D-Markers while enabling strong feature interactions. ***(ii)*** Despite their theoretical sophistication, attention-based mechanisms show surprisingly modest performance (average accuracy: 88.20%), indicating that the complementary nature of transformer and D-Marker features might be better captured through simpler, direct interactions rather than learned attention weights. ***(iii)*** The integration of Hadamard multiplication consistently enhances the performance of other fusion methods, as evidenced by the improvements in FiLM (90.33% → 93.00%), Bilinear Pooling (91.80% → 94.23%), and Factorized Bilinear Pooling

Table 4
Performance comparison of different feature fusion techniques across benchmark datasets. Results are reported in terms of classification accuracy (%). The **bold** values indicate the best performance for each dataset, while underlined values represent the second-best performance.

Fusion Method	Dataset				Average
	UvA-NEMO	MMI	SPOS	BBC	
Concatenation	87.9	**99.7**	97.2	95.0	94.95
Gated Concatenation	87.9	94.7	93.0	95.0	92.65
Additive Fusion	81.4	89.7	92.0	85.0	87.03
Hadamard (Multiplicative) Fusion	**88.7**	**99.7**	**98.5**	**100.0**	**96.73**
Gated Concatenation + Hadamard	87.9	**99.7**	93.4	**100.0**	95.25
Attention	87.1	89.4	91.6	90.0	89.53
Multi-head Attention	85.4	91.7	90.4	85.0	88.13
Cross Attention	83.8	92.3	85.7	90.0	87.95
Multi-head Cross Attention	85.4	92.1	86.2	85.0	87.18
Feature-wise linear modulation (FiLM)	83.8	94.7	92.8	90.0	90.33
FiLM + Hadamard	87.1	92.1	97.8	95.0	93.00
Bilinear Pooling	83.8	95.6	92.8	95.0	91.80
Bilinear Pooling + Hadamard	87.1	93.1	96.7	**100.0**	94.23
Factorized Bilinear Pooling	83.8	93.3	97.7	95.0	92.45
Factorized Bilinear Pooling + Hadamard	85.4	94.7	95.4	**100.0**	93.88

(92.45% → 93.88%). ***(iv)*** Simple concatenation achieves remarkably competitive performance (94.95% average), particularly on the MMI dataset (99.7%), suggesting that allowing the network to learn feature interactions through subsequent layers can be highly effective. ***(v)*** Higher-order interaction methods, while theoretically more expressive, show diminishing returns compared to simpler approaches, with factorized variants performing comparably to their full counterparts while being computationally more efficient. ***(vi)*** The effectiveness of different fusion strategies exhibits dataset-dependent variations, with multiple methods achieving perfect accuracy on the BBC dataset, while the UvA-NEMO dataset presents a more challenging scenario with lower overall performance across all methods.

4.3. Comparison with state-of-the-art

Our Hadamard fusion approach demonstrates consistent improvements over existing methods across all benchmark datasets. The following can be observed when comparing the performance of our method with existing approaches: ***(i)*** We achieve state-of-the-art performance on three out of four datasets, with particularly significant gains on SPOS (98.5%) and BBC (100%), surpassing both traditional feature-based approaches and recent deep learning methods. ***(ii)*** While Wu et al. [18] maintains the lead on UvA-NEMO (93.9% vs. our 88.7%), their method relies on manual landmark initialization and extensive preprocessing, whereas our approach is fully automatic and end-to-end trainable. ***(iii)*** Notably, we improve upon DeepMarkerNet [28], the previous state-of-the-art that also utilizes D-Marker information but through multi-task learning rather than direct feature fusion. Our method shows consistent gains across all datasets (UvA-NEMO: +0.8%, MMI: +0.7%, SPOS: +1.3%, BBC: +5.0%), demonstrating the superiority of explicit feature interaction over auxiliary task supervision.

The effectiveness of our Hadamard fusion strategy is visually demonstrated through t-SNE visualizations in Fig. 4. From the figure, we observe the following: ***(i)*** The baseline approach without D-Marker features shows significant overlap between spontaneous and posed smiles, indicating limited discriminative power. ***(ii)*** While DeepMarkerNet's multi-task learning approach improves feature separability, there is room for improvement and the decision boundaries are still coverlapping. ***(iii)*** In contrast, our Hadamard fusion method achieves remarkably clear class separation, suggesting that direct multiplicative interactions between transformer and D-Marker features create more discriminative representations compared to indirect supervision through auxiliary tasks. This is particularly evident in the distinct clustering of spontaneous

Table 5
Performance comparison with state-of-the-art methods on spontaneous smile recognition. Results reported as classification accuracy (%). Deep learning-based methods are highlighted in gray. Best results are in **bold**, second-best are underlined.

Method	UvA-NEMO	MMI	SPOS	BBC
Cohn'04 [13]	77.3	81.0	73.0	75.0
Dibeklioglu'10 [14]	71.1	74.0	68.0	85.0
Pfister'11 [15]	73.1	81.0	67.5	70.0
Wu'14 [16]	91.4	86.1	79.5	90.0
Dibeklioglu'15 [17]	89.8	88.1	77.5	90.0
Wu'17 [18]	**93.9**	92.2	81.2	90.0
Mandal'17 [19]	80.4	-	-	-
Mandal'16 [22]	78.1	-	-	-
RealSmileNet'20 [23]	82.1	92.0	86.2	90.0
PSTNet'22 [25]	72.9	94.3	87.1	95.0
P4Transformer'21 [26]	74.9	91.3	82.9	85.0
Vanilla ViT'20 [27]	78.4	99.0	93.5	95.0
MeshSmileNet'22 [24]	85.0	99.0	94.4	95.0
DeepMarkerNet'24 [28]	87.9	**99.7**	97.8	95.0
Ours (Hadamard Fusion)	88.7	**99.7**	**98.5**	**100.0**

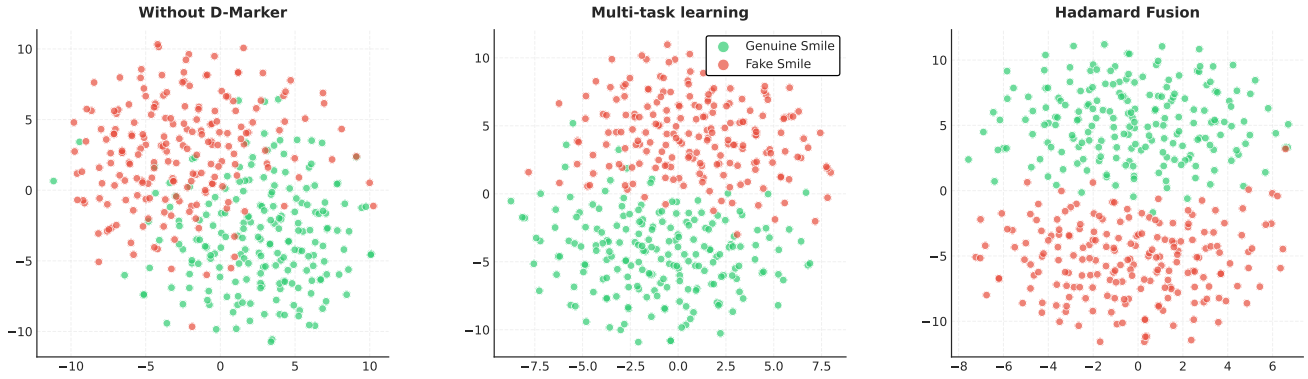

Figure 4: t-SNE visualization of learned feature embeddings from different approaches on the UvA-NEMO dataset. Left: baseline without D-Marker features shows significant class overlap. Middle: DeepMarkerNet's multi-task learning approach [28] demonstrates improved but still overlapping separation. Right: our Hadamard fusion method achieves better class separation with clearer decision boundaries.

and posed smile features with less overlaps and distinct decision boundaries, validating our hypothesis that explicit feature fusion more effectively leverages the complementary nature of learned and handcrafted features.

5. Discussion

Computational Efficiency. Our Hadamard fusion approach offers significant computational advantages over the multi-task learning framework (DeepMarkerNet). While DeepMarkerNet requires extensive hyperparameter tuning to balance the weighted loss function between smile classification (binary cross-entropy) and D-Marker prediction (MSE), our method eliminates this complexity entirely. The multi-task approach necessitates careful calibration of the α parameter that weights these losses, often requiring multiple training runs and validation experiments to find optimal values. In contrast, our direct feature fusion strategy requires no such balancing, reducing both training time and computational resources. Additionally, our method achieves superior performance without the overhead of D-Marker prediction during training, leading to more efficient optimization dynamics and faster convergence.

Generalization Analysis. The enhanced performance of our fusion-based approach across datasets suggests that multiplicative feature interactions effectively capture universal patterns distinguishing genuine from posed smiles. This is particularly evident in our strong results on the challenging SPOS dataset (98.5%) and perfect accuracy on BBC (100%), despite these datasets' varying recording conditions and resolutions. The success can be attributed to two key factors: (i) the Hadamard fusion's ability to preserve spatial correspondence between transformer and D-Marker features while enabling rich feature interactions, and (ii) the complementary nature of learned and handcrafted features, where transformer features capture complex spatiotemporal patterns while D-Markers provide physiologically-grounded priors.

Feature Interaction Analysis. The clear separation between genuine and posed smile features in our t-SNE visualizations provides insights into why Hadamard fusion outperforms alternative approaches. Unlike multi-task learning, which influences feature learning indirectly through loss functions, direct multiplicative fusion creates an explicit bottleneck where features must be mutually informative to contribute to the final prediction. This encourages the transformer to learn features that complement D-Marker information, rather than potentially redundant or conflicting representations. The effectiveness of this approach is further evidenced by the consistent improvement observed when Hadamard fusion is combined with other techniques (FiLM, bilinear pooling), suggesting its fundamental value in leveraging complementary feature types.

Limitations and Future Work. While our method demonstrates strong performance across benchmark datasets, several important directions remain for future investigation. ***(i)*** Although we achieve state-of-the-art results in within-dataset evaluations, comprehensive cross-dataset experiments (training on one dataset and testing on others) would provide deeper insights into the framework's generalization capabilities across different recording conditions and demographics. ***(ii)*** Our current implementation builds upon MeshSmileNet's transformer architecture; exploring integration with other state-of-the-art transformer variants like RealSmileNet could potentially yield further performance improvements through their unique architectural innovations. ***(iii)*** While we extensively evaluated different fusion strategies, systematic ablation studies on other architectural components (number of layers in classifier, hyperparameter sensitivity, etc.) could reveal additional optimization opportunities. ***(iv)*** From an application perspective, developing a web-based API for real-time smile authenticity analysis would facilitate broader practical adoption and enable real-world validation of our approach. These limitations present clear pathways for future research, particularly in enhancing model robustness and bridging the gap between academic research and practical deployment.

6. Conclusion

In this paper, we present a novel feature fusion framework, D-FuseSmileNet, for spontaneous smile recognition that effectively combines transformer-based features with physiologically-grounded D-Markers. Through extensive experimentation with 15 different fusion strategies, we demonstrate that Hadamard multiplicative fusion achieves optimal performance by preserving spatial correspondence while enabling rich feature interactions. Our approach surpasses state-of-the-art performance across multiple benchmark datasets, and across all deep-learning based methods, achieving 88.7% accuracy on UvA-NEMO, 99.7% on MMI, 98.5% on SPOS, and 100% on BBC. The method's effectiveness is particularly evident in the clear separation of feature embeddings

visualized through t-SNE, suggesting that direct multiplicative interactions better leverage the complementary nature of learned and handcrafted features compared to previous multi-task learning approaches. Moreover, our framework offers practical advantages through reduced computational complexity and elimination of loss-balancing hyperparameters. These findings provide valuable insights for future research in multimodal feature fusion for facial expression analysis and affective computing.

References

[1] Z. Dong, G. Wang, S. Lu, L. Dai, S. Huang, Y. Liu, Intentional-deception detection based on facial muscle movements in an interactive social context, Pattern Recognition Letters 164 (2022) 30–39.

[2] P. H. Barrett, The Works of Charles Darwin: Vol 23: The Expression of the Emotions in Man and Animals, Routledge, 2016.

[3] P. Ekman, R. J. Davidson, W. V. Friesen, The duchenne smile: Emotional expression and brain physiology: Ii., Journal of personality and social psychology 58 (1990) 342.

[4] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, J. Kissler, Mapping the emotional face. how individual face parts contribute to successful emotion recognition, PloS one 12 (2017) e0177239.

[5] D. Sarma, M. K. Bhuyan, Methods, databases and recent advancement of vision-based hand gesture recognition for hci systems: A review, SN Computer Science 2 (2021) 436.

[6] N. D. Bruce, C. Wloka, N. Frosst, S. Rahman, J. K. Tsotsos, On computational modeling of visual saliency: Examining what's right, and what's left, Vision Research 116 (2015) 95–112. Computational Models of Visual Attention.

[7] S. Y. Oh, J. Bailenson, N. Krämer, B. Li, Let the avatar brighten your smile: Effects of enhancing facial expressions in virtual environments, PloS one 11 (2016) e0161794.

[8] K. Lander, N. L. Butcher, Recognizing genuine from posed facial expressions: exploring the role of dynamic information and face familiarity, Frontiers in Psychology 11 (2020) 1378.

[9] S. Li, W. Deng, Deep facial expression recognition: A survey, IEEE transactions on affective computing 13 (2020) 1195–1215.

[10] M. Kawulok, J. Nalepa, J. Kawulok, B. Smolka, Dynamics of facial actions for assessing smile genuineness, Plos one 16 (2021) e0244647.

[11] O. A. Hassen, N. A. Abu, Z. Z. Abidin, S. M. Darwish, A new descriptor for smile classification based on cascade classifier in unconstrained scenarios, Symmetry 13 (2021) 805.

[12] D. E. Ratnawati, S. Anam, et al., Features selection for classification of smiles codes based on their function, in: ISRITI, IEEE, 2019, pp. 103–108.

[13] J. F. Cohn, K. L. Schmidt, The timing of facial motion in posed and spontaneous smiles, International Journal of Wavelets, Multiresolution and Information Processing 2 (2004) 121–132.

[14] H. Dibeklioglu, R. Valenti, A. A. Salah, T. Gevers, Eyes do not lie: Spontaneous versus posed smiles, in: ACM Multimedia, 2010, pp. 703–706.

[15] T. Pfister, X. Li, G. Zhao, M. Pietikäinen, Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework, in: ICCV Workshops, IEEE, 2011, pp. 868–875.

[16] P. Wu, H. Liu, X. Zhang, Spontaneous versus posed smile recognition using discriminative local spatial-temporal descriptors, in: ICASSP, IEEE, 2014, pp. 1240–1244.

[17] H. Dibeklioğlu, A. A. Salah, T. Gevers, Recognition of genuine smiles, IEEE Transactions on Multimedia 17 (2015) 279–294.

[18] P.-p. Wu, H. Liu, X.-w. Zhang, Y. Gao, Spontaneous versus posed smile recognition via region-specific texture descriptor and geometric facial dynamics, Frontiers of Information Technology & Electronic Engineering 18 (2017) 955–967.

[19] B. Mandal, N. Ouarti, Spontaneous versus posed smiles—can we tell the difference?, in: CVIP 2016, Volume 2, Springer, 2017, pp. 261–271.

[20] P. Ekman, W. V. Friesen, Facial action coding system, Environmental Psychology & Nonverbal Behavior (1978).

[21] P. Ekman, Facial expression and emotion., American psychologist 48 (1993) 384.

[22] B. Mandal, D. Lee, N. Ouarti, Distinguishing posed and spontaneous smiles by facial dynamics, in: ACCV, Springer, 2017, pp. 552–566.

[23] Y. Yang, M. Z. Hossain, T. Gedeon, S. Rahman, Realsmilenet: A deep end-to-end network for spontaneous and posed smile recognition, in: H. Ishikawa, C.-L. Liu, T. Pajdla, J. Shi (Eds.), Computer Vision – ACCV 2020, Springer International Publishing, Cham, 2021, pp. 21–37.

[24] M. T. Faroque, Y. Yang, M. Z. Hossain, S. M. Naim, N. Mohammed, S. Rahman, Less is more: Facial landmarks can recognize a spontaneous smile, arXiv preprint arXiv:2210.04240 (2022).

[25] H. Fan, X. Yu, Y. Ding, Y. Yang, M. Kankanhalli, Pstnet: Point spatio-temporal convolution on point cloud sequences, arXiv preprint arXiv:2205.13713 (2022).

[26] H. Fan, Y. Yang, M. Kankanhalli, Point 4d transformer networks for spatio-temporal modeling in point cloud videos, in: CVPR, 2021, pp. 14204–14213.

[27] 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).

[28] M. J. Hasan, K. Rafat, F. Rahman, N. Mohammed, S. Rahman, Deepmarkernet: Leveraging supervision from the duchenne marker for spontaneous smile recognition, Pattern Recognition Letters 186 (2024) 148–155.

[29] I. Grishchenko, A. Ablavatski, Y. Kartynnik, K. Raveendran, M. Grundmann, Attention mesh: High-fidelity face mesh prediction in real-time, arXiv preprint arXiv:2006.10962 (2020).

[30] K. L. Schmidt, J. F. Cohn, Y. Tian, Signal characteristics of spontaneous facial expressions: Automatic movement in solitary and social smiles, Biological psychology 65 (2003) 49–66.