3DCV Final Project
Group 4 - 3D Mesh From Video

R13922169 Chin-Hung, Liu, R14521607 Jing-Zhe, Lin, and R14921010 Pin-Wei, Huang

National Taiwan University

Abstract. High-fidelity 3D facial mesh reconstruction from monocular video is essential for realistic digital avatars, yet many methods struggle with temporal flickering and inconsistent geometry. This paper proposes an enhanced framework based on SMIRK that integrates explicit temporal constraints into the FLAME parametric model. We introduce a temporal smoothness loss for expression and pose, alongside an identity consistency loss to prevent shape drift across frames. To evaluate our approach, we train a downstream MLP (EmotionVA) to predict valence, arousal, and expression categories from the extracted FLAME parameters. Experimental results on the AffectNet dataset demonstrate that our temporal regularization significantly improves the stability of facial representations, notably increasing the Concordance Correlation Coefficient (CCC) for valence prediction from 0.4866 to 0.5266. Our work provides a robust solution for generating smooth, identity-consistent 3D facial animations from video sequences. For our source code, demo video and more, please visit our github repository: `https://github.com/steveliu2000/3dcv2025_final`.

Keywords: 3D Facial Reconstruction · FLAME Model · Temporal Consistency · Monocular Video · Emotion Recognition.

1 Motivation

High-fidelity 3D facial mesh reconstruction from monocular images or videos has emerged as a pivotal research area in computer vision. Despite significant advancements, many existing methods predominantly treat video frames as independent images or fail to fully exploit the inherent temporal dependencies across frames. This often leads to temporal flickering and inconsistent geometry in the reconstructed sequences.

We hypothesize that integrating explicit temporal constraints can substantially enhance the reconstruction quality and stability. Rather than regressing high-dimensional vertex positions directly, modern frameworks often leverage 3D Morphable Models (3DMMs), such as the Basel Face Model (BFM) [5] and FLAME [3], to constrain the solution space within a low-dimensional, biologically plausible manifold. In this work, we propose to enforce temporal consistency on these parametric representations to achieve smoother and more accurate facial animations.

2 Related Work

2.1 FLAME: A Parametric Head Model

FLAME [3] is a state-of-the-art expressive 3D head model trained from extensive 4D scans. It provides a compact and controllable representation of facial geometry by decomposing it into linear subspaces of identity shape, expression, and pose. Formally, the FLAME model is defined as:

$$FLAME(\beta, \psi, \theta) \rightarrow (V, F) \tag{1}$$

where

- $\beta \in \mathbb{R}^{|\beta|}$ represents the identity shape parameters.
- $\psi \in \mathbb{R}^{|\psi|}$ represents the facial expression parameters.
- $\theta \in \mathbb{R}^{|\theta|}$ denotes the pose parameters (including neck, jaw, and eyeballs).

The model outputs a mesh with $V \in \mathbb{R}^{5023\times3}$ vertices and $F \in \mathbb{R}^{9976\times3}$ fixed triangulated faces.

2.2 Monocular face reconstruction

EMOCA [1] focuses on capturing the emotional content of facial expressions. It employs a convolutional encoder to regress FLAME parameters and utilizes an albedo model to extract detailed texture features. The reconstruction is guided by a multi-level supervision strategy, incorporating 2D landmarks, 3D geometry consistency, and specialized emotion-related losses. This allows EMOCA to recover fine-grained, semantically meaningful expressions that are often overlooked by standard geometric losses.

SMIRK [6] introduces a novel "analysis-by-neural-synthesis" approach. It decouples identity from motion by predicting expression and pose through a lightweight encoder, then reconstructs the input image using an image-to-image neural generator. A key contribution of SMIRK is the augmented cycle pass, which enhances the model's robustness and ability to reconstruct rare or extreme facial expressions by leveraging synthetic data in a self-supervised loop. See 3.2 for more details about SMIRK.

While the aforementioned methods achieve impressive results on static images, they do not inherently account for the temporal dynamics in video. Although some video-based methods like Neural Head Avatars [2] incorporate temporal information, they primarily focus on optimizing static, motion-unrelated features (such as identity or neutral geometry). In contrast, our work aims to bridge this gap by enforcing temporal constraints on both static (identity) and dynamic (expression and pose) features, ensuring holistic consistency across the temporal domain.

3 Method

3.1 Problem Formulation

Given a short video clip consisting of T consecutive frames $\{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_T\}$ depicting the same subject, our goal is to reconstruct a temporally consistent sequence of 3D facial meshes.

Instead of directly predicting dense vertex coordinates, we adopt the FLAME parametric model, where each frame $\mathbf{I}_t$ is mapped to a set of FLAME parameters:

$$(\beta_t, \psi_t, \theta_t) = f(\mathbf{I}_t), \tag{2}$$

While conventional monocular reconstruction methods process each frame independently, this formulation often leads to temporal jitter and identity drift when applied to videos. To address this issue, we explicitly enforce temporal consistency on the predicted FLAME parameters during training.

3.2 Baseline: SMIRK

Our method is built upon SMIRK [6]. It employs a convolutional encoder to regress FLAME parameters from a single image. Specifically, these parameters (β, ψ, θ) are fed into the FLAME model to generate a 3D facial mesh. This mesh is then processed through a differentiable renderer to produce a rendered image of the face. Subsequently, the original input image has the facial area masked out and is concatenated with the rendered mesh image. These combined inputs are fed into an image-to-image neural generator to reconstruct the original image.

SMIRK is trained using image-level supervision, including landmark losses, reconstruction losses, and perceptual losses, and has demonstrated strong performance on reconstructing detailed facial expressions from static images. However, the original SMIRK framework treats each input independently and does not model temporal dependencies, making it suboptimal for video-based reconstruction scenarios.

3.3 Temporal Regularization on FLAME Parameters

To extend SMIRK to video inputs, we introduce temporal regularization terms applied directly to the predicted FLAME parameters. Importantly, this is achieved without altering the original network architecture; temporal information is incorporated solely through the training objective.

Temporal Smoothness Loss. Facial expression and pose parameters are expected to vary smoothly over time. We therefore penalize abrupt changes between consecutive frames using a temporal smoothness loss:

$$\mathcal{L}_{smooth} = \frac{1}{T-1} \sum_{t=1}^{T-1} \left\| \mathbf{p}_{t+1} - \mathbf{p}_t \right\|_1, \tag{3}$$

where $\mathbf{p}_t$ denotes time-varying parameters such as expression, pose, jaw, or eyelid parameters at frame t. This loss effectively reduces temporal flickering in reconstructed facial motions.

Identity Consistency Loss. In contrast to expression and pose, the shape parameters β represent subject identity and should remain constant throughout a video. To enforce identity consistency, we minimize the variance of shape parameters across the entire clip:

$$\mathcal{L}_{id} = \frac{1}{T} \sum_{t=1}^{T} \left\| \beta_t - \bar{\beta} \right\|_2^2, \tag{4}$$

where $\bar{\beta}$ is the mean shape parameter across all frames. This formulation explicitly prevents identity drift that may occur when only local smoothness constraints are applied.

Overall Objective. The final training objective combines the original SMIRK losses with the proposed temporal regularization terms:

$$\mathcal{L} = \mathcal{L}_{SMIRK} + \lambda_{smooth} \mathcal{L}_{smooth} + \lambda_{id} \mathcal{L}_{id}, \tag{5}$$

where λ_{smooth} and λ_{id} control the relative importance of each temporal constraint.

4 Evaluation

4.1 Dataset: AffectNet

We conduct our evaluation on AffectNet [4], a large-scale facial emotion dataset widely used for both categorical and dimensional emotion analysis. Each sample in AffectNet provides valence, arousal, and expression label.

Valence and arousal . Valence is a continuous value ranging from -1 to 1, indicating emotional positivity (-1: negative, 1: positive). Arousal is a continuous value ranging from 0 to 1, representing emotional intensity (0: calm, 1: excited), as shown in Fig. 1.

Expression label. Each sample has an expression label corresponding to one of the 8 expression categories, as shown in Table 1.

Table 1. Labels and the corresponding expressions (adapted from [4]).

Label	1	2	3	4	5	6	7	8
Expression	Neutral	Happiness	Sadness	Surprise	Fear	Disgust	Anger	Contempt

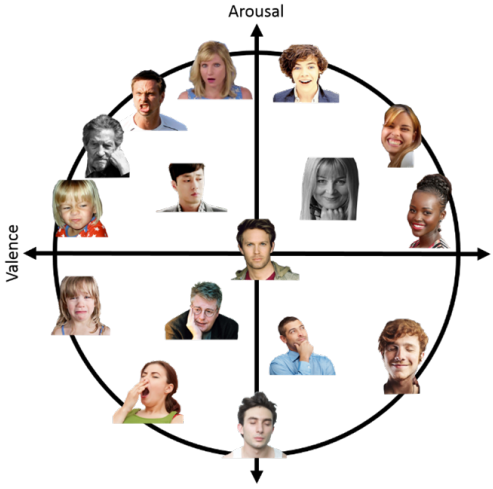

Fig. 1. Valence and arousal (taken from [4]).

4.2 Evaluation Framework

To evaluate the correctness of our method, we use root mean square error (RSME), concordance correlation coefficients (CCC) of the valence and arousal values, along with the accuracy of the expression classification as the evaluation metrics, which are also used in the original SMIRK paper.

However, we cannot compute the valence and arousal from the FLAME parameters directly. Therefore, we apply the same method as in [1]. First train an MLP, which uses expression parameters as inputs and outputs valence, arousal and expression categories, then we use this MLP to evaluate model, as shown in Fig. 2.

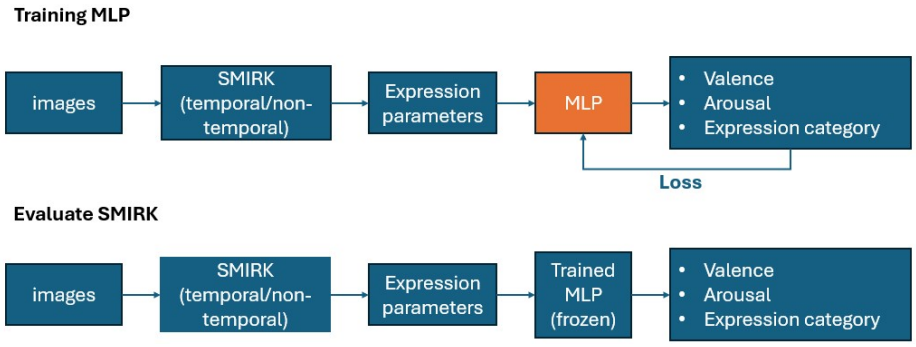

Fig. 2. The pipeline of the training and evaluation processes.

4.3 Results

The emotion recognition results of the two models, trained with temporal loss and non-temporal loss, respectively, are shown in Table 2. We report concordance correlation coefficient (CCC) and root mean square error (RMSE) of valence (V-) and arousal (A-), and expression classification accuracy (E-ACC).

$$RMSE(Y, \hat{Y}) = \sqrt{\mathbb{E}\left[(Y - \hat{Y})^2\right]} \tag{6}$$

$$CCC(Y, \hat{Y}) = \frac{2\sigma_Y \sigma_{\hat{Y}} \, \mathrm{PCC}(Y, \hat{Y})}{\sigma_Y^2 + \sigma_{\hat{Y}}^2 + (\mu_Y - \mu_{\hat{Y}})^2} \tag{7}$$

$$PCC(Y, \hat{Y}) = \frac{\mathbb{E}\left[(Y - \mu_Y)(\hat{Y} - \mu_{\hat{Y}})\right]}{\sigma_Y \sigma_{\hat{Y}}} \tag{8}$$

Table 2. Comparison between non-temporal and temporal SMIRK for emotion recognition

	V-CCC↑	V-RSME↓	A-CCC↑	A-RSME↓	E-ACC↑
Non-temporal	0.4866	0.3955	0.3951	**0.3576**	**0.3591**
Temporal	**0.5266**	**0.3811**	**0.3999**	0.36	0.3571

As shown in Table 2, temporal modeling does not significantly affect expression classification accuracy, with both models achieving similar performance (~ 0.36).

However, the temporal model representation yields a clear improvement in valence prediction, reducing RMSE from 0.3955 to 0.3811 and increasing CCC from 0.4866 to 0.5266. For arousal prediction, the temporal model achieves a slightly higher CCC, while RMSE remains comparable.

5 Discussion

Our experimental results demonstrate that incorporating temporal constraints during training enhances the representational power of the FLAME parameters, particularly for dimensional emotion metrics like valence. The improvement in Valence-CCC suggests that enforcing smoothness across frames helps the model capture more reliable emotional trajectories, reducing the noise typically found in per-frame estimation.

In contrast, discrete expression classification appears less sensitive to temporal context, likely because expression categories are dominated by instantaneous facial configurations rather than temporal dynamics.

Overall, these findings demonstrate that temporal SMIRK representations provide more informative features for valence-arousal regression tasks, while maintaining comparable performance for expression classification.

References

1. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20311–20322 (2022)
2. Grassal, P.W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., Thies, J.: Neural head avatars from monocular rgb videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18653–18664 (2022)
3. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
4. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
5. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. pp. 296–301. Ieee (2009)
6. Retsinas, G., Filntisis, P.P., Danecek, R., Abrevaya, V.F., Roussos, A., Bolkart, T., Maragos, P.: Smirk: 3d facial expressions through analysis-by-neural-synthesis. arXiv preprint arXiv:2404.04104 (2024)