Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

PersonaGAN: Conditional Stacked GAN Framework for Emotion-Aware Avatar Generation using Wearable Physiological Data

[1]Ankur Sharma, [2]Ambigavathi Munusamy

[1]*M.Tech. Scholar, School of Computing, Indian Institute of Information Technology Una,*

[2]*Assistant Professor, School of Computing, Indian Institute of Information Technology Una*

ARTICLE INFO

Received: 25 May 2026

Revised: 02 July 2026

Accepted: 10 July 2026

ABSTRACT

Over the past few years, wearable devices like smartwatches have become central to the seamless tracking of physiological signals that reveal a person's emotional and physical state. Although earlier research has shown that it is possible to identify stress and emotional states from this kind of data, the results are usually simplified to abstract measures like class labels, confidence scores, or statistical plots that have very little interpretability or value for the end users' experience. Besides, many emotion detection tools rely on externally visible features, especially facial expressions, which may not accurately represent the real emotional states if the expressions are being regulated either deliberately or unintentionally. Conversely, emotional states inferred from physiological signals are more accurate representations of the true affect, however, this area has not been sufficiently explored in terms of intuitive and human-centered interactive visualization. Here is where PersonaGAN comes in, a conditional stacked adversarial framework that converts physiological signals obtained from wristbands into expressive visual avatars. Wrist data from the WESAD dataset are first processed, then normalized and encoded into emotion-aware conditioning vectors that depict internal affective responses without depending on facial cues. The proposed framework is a two-stage conditional stacked generative adversarial architecture with perceptual refinement that helps in semantically coherent and visually realistic avatar generation.

Index terms: Affective Computing, Avatar Generation, Generative Adversarial Networks, Physiological Signals.

INTRODUCTION

The widespread use of wrist-worn wearable devices has made it possible to continuously and discreetly monitor physiological signals using highly popular consumer technologies like smartwatches. These gadgets detect a variety of physiological signals, such as Electrodermal Activity (EDA), Blood Volume Pulse (BVP), skin temperature, and accelerometer signals, which have become increasingly important sources of information for affective computing and emotion recognition applications. Most of the time, existing methods use machine learning and deep learning techniques to categorize physiological signals into predefined emotional or stress-related groups. While these techniques often attain very high classification accuracy, their results are usually limited to numerical labels, confidence scores, or statistical representations, thus offering only a very narrow interpretation and user engagement potential.

Converting physiological signals into expressive visual representations is quite a difficult cross-domain mapping problem. Physiological measurements naturally carry implicit and non-observable affective responses, but avatar generation needs explicit visual and semantic structures that can represent emotional characteristics. So, establishing significant connections between these different modalities still seems to be a difficult issue in affective visual synthesis.Generative Adversarial Networks (GANs) are a new type of models that can understand complex data distributions and create very realistic images [1]. Conditional GANs can do even more as they add extra information that allows for controlled image generation [2]. On top of that, stacked GAN architectures take the quality of

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

generation even further by using coarse to fine refinement strategies [3,4]. Moreover, GAN-based methods have unlocked a great deal of potential in healthcare and biomedical areas by enabling data augmentation, privacy preservation, and representation learning [5,6]. While we have seen quite some advancement in wearable emotion recognition and generative modelling, these two areas have barely interacted with each other. Usually, physiological sensing systems deliver emotion labels or probabilities, and except the visual input, the existing avatar generation methods mainly use facial images, motion capture data, or depth information. Most of the existing methods for avatar synthesis are aimed at appearance transfer, facial animation, or motion-driven generation. As a result, physiological signals are not directly included as conditioning information. Therefore, present avatar systems frequently do not depict the user's fundamental emotional state that is deduced from physiological measurements.

In an effort to overcome this drawback, the paper introduces PersonaGAN, a health-aware conditional stacked generative adversarial framework that converts physiological signals obtained from a smartwatch into visual avatars that are emotion-aware. The suggested framework integrates normalized physiological features and affective labels into a single conditioning representation and uses a two-stage conditional GAN architecture to produce affinitive and visually perfect avatars. The first stage creates the overall avatar, whereas the second stage enhances the visual quality and structural consistency through stacked adversarial learning. The experimental results show that the framework proposed here is able to perform stable training, yield low Fréchet Inception Distance (FID) scores, and maintain a high level of conditional consistency across different emotion categories. The main contributions of this work is as follows:

- A conditional stacked GAN framework, focused on health, has been introduced. It aims at creating emotion-aware avatars based on physiological signals collected from smartwatches along with their corresponding affective labels.
- A novel two-step adversarial generation approach is proposed that significantly enhances the semantic consistency, the visual realism, and the structural coherence of the generated avatar.
- Through numerous and thorough quantitative and qualitative assessments, such as FID analysis, conditional consistency checks, and ablation experiments, the results clearly show that the fusion of physiological features with emotional labels is a potent approach for avatar creation.

The rest of this paper is structured as follows. Section II examines existing research in areas such as wearable emotion recognition, conditional generative modelling, and avatar generation. Section III introduces the new PersonaGAN framework. In Section IV, we analyze the experimental results and evaluation, while in Section V, we wrap up the paper.

I. RELATED WORK

With the increase in wearable device usage such as smartwatches and bio-sensing technology, research in wearable-based emotion and stress recognition has gained significant attention. The WESAD dataset is quite popular this domain since it includes synchronized recordings of physiological data from human subjects obtained from both heart and wrist sensors [7]. Even if sensors on the chest generally give better signals, for objective reasons their limited use in daily wear has led to the rising of wrist-only sensing. Machine learning and deep learning methods have all been tried on WESAD along with similar datasets. Convolutional neural networks have shown great results for classifying stress and emotions [8], while ensemble and multimodal learning methods are providing robustness and higher accuracy of predictions [9,10]. A number of works also indicate that physiological signals from the wrist, especially electrodermal activity and skin temperature, are still highly revealing without the need for chest sensors [11]. Despite such progress, the majority of present-day methods concentrate mainly on predictive classification and offer numerical labels or probabilities that are very limited in terms of understandable and engaging representation. Interactive emotion-aware systems that have been recently developed proved that visually understandable representations of emotional states can increase user involvement, therapeutic feedback, and human-centered interaction in healthcare and rehabilitation settings [12,13]. Nevertheless, these systems do not directly deal with the creation of physiologically influenced visual avatars based on wearable sensing data, thus leaving a large gap between wearable affect detection and natural emotion visualization.

At the same time, GANs have become incredibly effective tools for modeling complex data distributions and creating realistic images. Conditional GANs, for instance, enhance this idea by adding extra information that helps guide the

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

image creation process through semantic conditioning [2]. Additionally, auxiliary-classifier versions further enhance the semantic accuracy by jointly optimizing both the adversarial and classification objectives. Stacked GAN architectures like StackGAN and StackGAN++ break down the task of image generation into multiple stages, which allows for the coarse-to-fine refinement and also makes the training more stable when the conditioning information doesn't have a clear spatial structure [3,4]. These features make stacked conditional architectures especially fitting for conditioning signals that come from non-visual domains, such as physiological measurements. GAN-based approaches have already found their way into healthcare and affective computing for purposes such as data augmentation, modality translation, and privacy-preserving biomedical applications [14–16]. In the field of affective computing, emotion-aware emoji generation [17] and adversarial strategies for physiological emotion recognition [18] are examples of how generative modeling is expanding its capabilities. Nevertheless, these techniques usually work with specific emotion vectors or representations at the signal level and do not produce visually realistic avatars that are physiologically conditioned.

Over the last few years, avatar creation has also turned a significant corner. Initial techniques concentrated on the learning of parameterized or free-form avatar representations [19]. However, nowadays, the focus is on controllable 3D and 3D-aware avatar synthesis from a very limited number of visual inputs [20,21]. Methods for human reconstruction from a single image are also constantly proving that one can extract realistic geometry and appearance with minimal visual data [22,23].Top-tier systems like AvatarArtist and Snapmoji provide avatars that are not only highly expressive but also visually convincing, perfect for interactive applications [24,25].Further studies have delved into multimodal instruction-guided avatar animation [26], comprehensive avatar modeling pipelines [27], and geometry-aware reconstruction techniques that facilitate immersive digital-human interaction [28–30].Even with these dramatic breakthroughs, the majority of avatar creation tech stacks still predominantly depend on visual, facial, or motion-based inputs and don’t allow for the integration of physiological sensing or internal affective information as conditioning signals.

TABLE I
PERSONAGAN VS. RELATED METHODS

Method	**Physiological Conditioning**	**Avatar Generation**	**Multi-Stage GAN**
Emoji-based GAN [17]	No	No	No
GANser [18]	Yes	No	No
AvatarArtist [24]	No	Yes	No
Snapmoji [25]	No	Yes	No
PersonaGAN	Yes	Yes	Yes

The collection of works reviewed in the article exhibit three major shortcomings. First, the wearable emotion recognition systems are mainly designed for predictive classification. As a result, their outputs lack interpretability. Second, most of the avatar creation methods do not physiologically sense and therefore, cannot reflect the internal emotional changes. Third, despite the advances made in affective computing and avatar synthesis, the two are still largely being researched in their separate tracks. The repercussions of these three limitations are most severely felt in applications where internal physiological states are crucial but, at the same time, hard to directly observe, such as

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

in healthcare monitoring, stress management, telepresence communication, rehabilitation support, high-performance sports, industrial safety monitoring, disaster-response training, and mission-critical operational environments. In these cases, just numerical indicators are not likely to give the most intuitive or usable representations of the user's emotional state. These findings, in turn, call for the creation of models that would merge wearable physiological sensing and visually interpretable emotion-aware avatar generation.

II. PROPOSED PERSONAGAN FRAMEWORK

In all of our experiments, we used the PyTorch deep learning framework, and they were carried out on a workstation with an AMD Ryzen 9 7940HS processor, 16 GB RAM, and an NVIDIA GeForce RTX 4070 Laptop GPU with 8 GB VRAM. The training was done on a 64-bit Windows operating system with CUDA acceleration turned on for GPU-based optimization and inference. Here, we outline the PersonaGAN framework that we've developed to synthesize emotion-aware avatars from physiological signals obtained via a smartwatch. The framework is composed of four key steps: (i) capturing and preprocessing data using the wearable, (ii) creating the conditioning vector, (iii) generating the avatar coarsely conditioned, and (iv) improving the high-resolution avatar via a stacked generative adversarial network.

A. Wearable Data Collection and Preprocessing

The PersonaGAN uniquely relies on the wrist-only portion of the WESAD dataset [7] for its construction. Although WESAD includes physiological signals from both chest-mounted and wrist-mounted sensors, chest devices require specialized hardware and do not reflect the capabilities of consumer smartwatches. For this reason, only wrist-based signals were considered. The selected modalities include blood volume pulse, electrodermal activity, skin temperature, and tri-axial accelerometer data. The original signals, sampled at 32 Hz, were resampled to 1 Hz using window-based aggregation to align them with the available emotion labels. Data segments corresponding to undefined or transition states were excluded, and the remaining samples were grouped into four emotion classes: Baseline, Amusement, Meditation, and Stress. To address class imbalance, controlled oversampling was applied during preprocessing. Further, the wrist-only sub set features are selected to realistically model consumer smart watch sensing. All these features are normalized prior to conditioning vector construction, ensuring consistent feature scaling during adversarial training, as listed in Table 2.

TABLE II WRIST SUBSET FEATURES FOR PERSONAGAN CONDITIONING

Feature	**Rate**	**Range**
BVP (Wrist_BVP_0)	1 Hz	[-1773.78, 1789.09] a.u.
EDA (Wrist_EDA_0)	1 Hz	[-0.089, 9.862] µS
Temp (Wrist_TEMP_0)	1 Hz	[29.40,35.98] ◦C
ACC X (Wrist_ACC_0)	1 Hz	[-158.60, 174.26] g
ACC Y (Wrist_ACC_1)	1 Hz	[-149.48, 140.38] g
ACC Z (Wrist_ACC_2)	1 Hz	[-188.55, 180.17] g

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

B. Conditioning Vector Construction

All wrist-derived physiological features are first normalized to the range [0,1] using Min–Max scaling to ensure numerical stability during training. The emotion labels were then encoded using a one-hot representation and concatenated with the normalized features to form a fixed-length conditioning vector as represented in (1):

$$c = [f_{\{norm\}}, y_{\{one-hot\}}] \quad (1)$$

Here, $\mathbf{f_{norm}}$ denotes the normalized wrist-derived physiological features, and $\mathbf{y_{one\text{-}hot}}$ represents the corresponding emotion label. In order to assess how well wrist-only signals can be distinguished, a gradient-boosted decision tree (XGBoost) classifier is trained and tested with the same normalized feature representation that was used for PersonaGAN conditioning. On the balanced dataset, the classifier achieved strong classification performance, consistent with results reported in [9, 10]. Subsequently, a lightweight Convolutional Neural Network (CNN) trained on the same features set showed similar performance. These models served only for validation and analysis purposes and are not used during adversarial training. The predicted results suggest that normalized wrist-derived physiological features contain enough emotional information for conditional avatar generation. To prevent the generator from relying exclusively on categorical emotion supervision, PersonaGAN jointly conditions the synthesis process on both continuous physiological embeddings and high-level affective labels.

The physiological feature vectors preserve continuous biological variations extracted from multimodal smartwatch signals, whereas the predicted emotion labels provide coarse semantic guidance for affect-aware avatar generation. To further analyze the relative contribution of physiological signals and categorical labels, additional ablation experiments were conducted using: (i) labels-only conditioning, (ii) physiology-only conditioning, and (iii) combined physiological and affective conditioning. The results demonstrate that physiology-only conditioning produces structurally unstable avatars, whereas labels-only conditioning preserves 5 coarse semantic structure but lacks physiological modulation. The integrated conditioning approach delivered the most semantically coherent results combined with the lowest FID score. This means that the generator is not only relying on the categorical labels but gaining additional physiological information when synthesizing the avatars. At inference time, the trained XGBoost can optionally infer emotion labels from incoming smartwatch data. The inferred labels are then combined with the normalized physiological features and used as input to the conditional generator for live avatar synthesis.

C. Avatar Dataset Construction

To the best of our knowledge, there is no public dataset that directly connects physiological signals with avatar images. Therefore, we created a custom avatar dataset using the Dice Bear Avatars API [31]. Since the DiceBear API does not provide deterministic affective control or explicit emotion-specific generation seeds, avatar images were initially generated using randomized combinations of facial attributes, hairstyles, accessories, background settings, and color palettes. The resulting avatars were manually inspected and categorized into the predefined emotion classes, namely baseline, amusement, meditation, and stress, based on visually perceived emotional characteristics and expression styles.

The initial generation process produced approximately 106 stress, 74 baseline, 69 amusement, and 51 meditation avatar

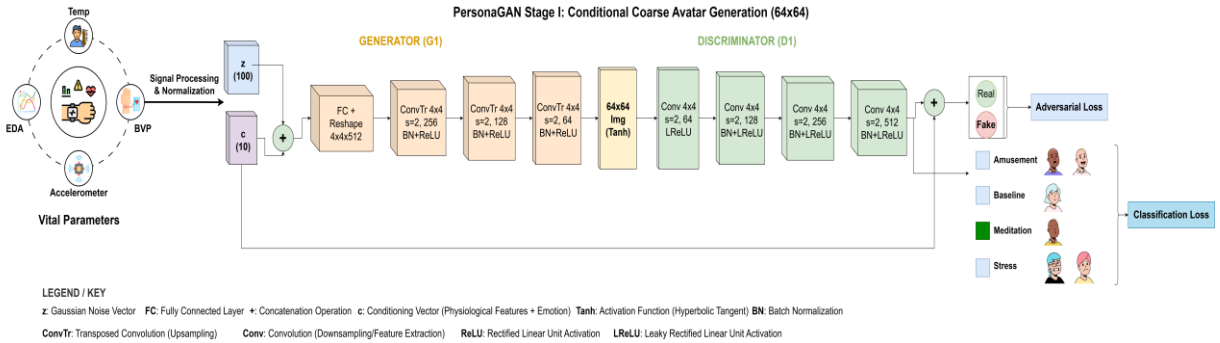


Fig. 1 Stage 1 Conditional coarse avatar generation architecture.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

samples. To maintain balanced class distributions during GAN training, each emotion category was subsequently expanded to 300 samples through controlled duplication, upsampling, and appearance-preserving variations. Samples that were visually corrupted, duplicated, or semantically inconsistent were excluded through manual verification in order to enhance dataset quality and ensure consistency in diversity. The images that were generated were subsequently sorted into emotion-specific folders, thus facilitating the use of conditional supervision during adversarial training. To maintain the same quality among the images, every avatar picture was reset to a predetermined resolution and color space, whereas the backgrounds got a uniform color in order to eliminate any unintentional visual cues during training.

D. Stage 1: Conditional Coarse Avatar Generation

The first-stage architecture of PersonaGAN is shown in Fig. 1. This architecture is used to perform conditional low-resolution avatar synthesis through an auxiliary classifier GAN formulation. The generator G_1, represented in (2), takes as input a concatenation of a 100 dimensional Gaussian noise vector $\mathbf{z}$ and a 10-dimensional conditioning vector $\mathbf{c}$. The conditioning vector encodes normalized wrist-based physiological features along with a four-class one-hot emotion representation. This joint latent embedding is first projected through a fully connected layer and then passed through a sequence of upsampling convolutional (ConvTranspose) blocks with batch normalization and ReLU activations, producing a coarse 64 × 64 RGB avatar image:

$$\hat{x}_{64} = G_{1(z,c)} \tag{2}$$

The corresponding discriminator, denoted as D_1, follows a design similar to Auxiliary Classifier GAN (AC-GAN) with two output branches: (i) an adversarial head that distinguishes real images from generated ones, and (ii) an auxiliary classification head that predicts the associated emotion class. This dual head structure explicitly enforces alignment between the generated avatars and the underlying physiological emotion labels [2, 32]. In addition, the adversarial loss drives the generator to match the distribution of real avatar images and is expressed in (3):

$$\mathcal{L}_{adv}^{(1)} = E_{x,c}[\log D_1(x,c)] + E_{z,c}[\log(1 - D_1(G_1(z,c),c))] \tag{3}$$

Similarly, the auxiliary classification loss enforces semantic consistency between the generated avatars and the target emotion class and is given in (4):

$$\mathcal{L}_{cls}^{(1)} = \mathbb{E}_{x,y}[-\log D_1^{cls}(y|x)] + \mathbb{E}_{z,y}[-\log D_1^{cls}(y|G_1(z,c))] \tag{4}$$

The overall optimization objective for Stage 1 is formulated by (5):

$$\mathcal{L}_{Stage1} = \mathcal{L}_{adv}^{(1)} + \lambda_{cls}\mathcal{L}_{cls}^{(1)} \tag{5}$$

Where, λ_{cls} controls the weight of the auxiliary classification loss. Empirically, training converges after approximately 6000 epochs, producing coarse avatars that reliably reflect the target emotions. Stage 1 training was implemented using the PyTorch deep learning framework and optimized using the Adam optimizer with a learning rate of 2×10^{-4} and momentum parameters $\beta1 = 0.0$ and $\beta2 = 0.9$. The batch size was fixed at 128 during adversarial training.

E. Stage 2: Conditional Refinement Using Stacked GAN

The stage 2 performs high-resolution refinement of the coarse avatars generated in stage 1 using a conditional stacked GAN strategy inspired by StackGAN and StackGAN++ [3, 4]. The objective of this stage is to enhance visual fidelity, texture sharpness, and structural coherence while preserving the emotion semantics encoded during stage 1. As illustrated in Fig. 2, the pretrained stage 1 generator G_1 is frozen and used to produce low-resolution conditional avatars $\hat{\mathbf{x}}_{64}$, which serve as input to the refinement generator G_2. The stage 2 generator adopts an encoder–residual–decoder architecture and produces refined 128 × 128 RGB avatars as given in (6). The encoder–residual–decoder layout was chosen as it allows for a clean dissection of the steps at which features are compressed, semantics are refined, and the high resolution of the output is restored. The encoder gradually derives multi-scale structural representations of the coarse stage 1 avatar image, while residual blocks maintain semantic similarity and stabilize

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

the passage of features during the emotion-aware refinement. The decoder then resumes the high-frequency facial details and the sharpness of the texture through the upsampling process. Compared with direct image-to-image translation architectures, this design provided improved training stability, reduced visual artifacts, and better preservation of emotion-conditioned facial structures during high-resolution avatar synthesis. The refined avatar generation process is formulated as:

$$\hat{x}_{128} = G_2(\hat{x}_{64}, c) \qquad (6)$$

Here, **c** denotes the same smartwatch-derived conditioning vector used in stage 1. The encoder progressively downsamples the stage 1 output to extract multi-scale feature representations, after which the conditioning vector **c** is fused through channel wise broadcasting. A stack of residual blocks then refines the intermediate representations, and the decoder upsamples the features to generate a high-resolution avatar image. The corresponding stage 2 discriminator, D_2, follows the same dual-head auxillary classification design as D_1,

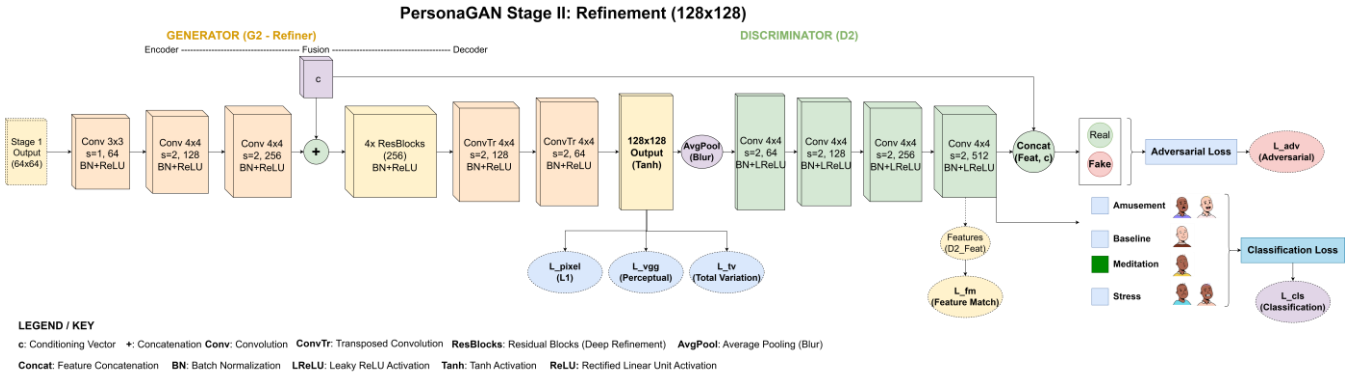


Fig. 2 Stage 2 Conditional refinement architecture.

consisting of an adversarial real/fake head and anauxiliary emotion classification head. In addition, D_2 exposes intermediate feature maps that are used to compute a feature-matching loss. To improve training stability and reduce sensitivity to high-frequency artifacts, alight blur preprocessing operation is applied to both real and generated images before discriminator evaluation. The stage 2 optimization objective integrates adversarial, reconstruction, and perceptual constraints and is expressed as given in(7):

$$\mathcal{L}_{Stage2} = \mathcal{L}_{adv}^{(2)} + \lambda_{cls}\mathcal{L}_{cls}^{(2)} + \alpha\mathcal{L}_{L1} + \beta\mathcal{L}_{FM} + \gamma\mathcal{L}_{perc} + \delta\mathcal{L}_{TV} \qquad (7)$$

Where, $\mathcal{L}_{adv}^{(2)}$ denotes the adversarial loss, $\mathcal{L}_{cls}^{(2)}$ controls the weight of the auxiliary classification loss, $\mathcal{L}_{L1}$ enforces pixel level consistency, $\mathcal{L}_{FM}$ aligns discriminator feature statistics, $\mathcal{L}_{perc}$ captures perceptual similarity using a pretrained network, and $\mathcal{L}_{TV}$ leads to spatial smoothness. Thus, refining stage achieves stable convergence and produces visually consistent, emotion-aware avatars with a Fréchet Inception Distance (FID) of approximately 2, and indicating high perceptual similarity between generated and reference avatar distributions. The refinement stage used Adam optimization with generator and discriminator learning rates of 2×10^{-4} and 1.5×10^{-4}, respectively. Using momentum parameters $\beta1=0.0$ and $\beta2=0.9$ with a batch size of 128. Multiple complementary objectives, including adversarial loss, perceptual loss, feature-matching loss, L1 reconstruction loss, and total variation regularization, were jointly optimized to improve visual realism, texture consistency, and structural stability during refinement.

F. Stage 2 Hybrid Loss Integration: To ensure the both visual realism and semantic correctness during high-resolution refinement, stage 2 is trained using a hybrid multi-objective loss formulations. This can balance adversarial supervision, emotion consistency, reconstruction fidelity, perceptual similarity, and spatial smoothness. Similar to (3), the adversarial loss of stage-2 ($\mathcal{L}_{adv}^{(2)}$) is formulated to encourage the generator to produce avatars that are indistinguishable from real samples as presented in(8):

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

$$\mathcal{L}_{adv}^{(2)} = \mathbb{E}_{x,c}[\log D_2(x,c)] + \mathbb{E}_{\hat{x}_{128},c}[\log(1 - D_2(\hat{x}_{128},c))] \quad (8)$$

Let $D_2^{cls}(\cdot)$ denote the classification head of the stage2 discriminator and y the ground-truth emotion label. An auxiliary classification loss is applied to both real and generated samples to enforce emotion-consistent generation and is defined in (9):

$$\mathcal{L}_{cls}^{(2)} = \mathbb{E}_{x,y}[-\log D_2^{cls}(y|x)] + \mathbb{E}_{\hat{x}_{128},y}[-\log D_2^{cls}(y|\hat{x}_{128})] \quad (9)$$

Let x denote the reference avatar image. The $\ell 1$ reconstruction loss ($\mathcal{L}_{L1}$) constrains the stage 2 generator to preserve low frequency structural content while limiting excessive deviation from the stage 1 output and is defined in (10):

$$\mathcal{L}_{L1} = \mathbb{E}[\|x - \hat{x}_{128}\|_1] \quad (10)$$

The feature-matching loss in (11) is computed using intermediate discriminator activations to stabilize adversarial training and suppress high-frequency artifacts. Here, $D_2^{(l)}(\cdot)$ denotes the feature map extracted from the l-th discriminator layer.

$$\mathcal{L}_{FM} = \sum_l \left\| D_2^{(l)}(x) - D_2^{(l)}(\hat{x}_{128}) \right\|_1 \quad (11)$$

Since pixel-wise losses fail to capture high-level visual similarity, a perceptual loss, defined in (12), is computed using features from a pretrained CNN, with $\phi_k(\cdot)$ representing the activation of the k-th network layer. In the implementation, a pretrained VGG16 network is employed to extract intermediate perceptual feature representations from multiple convolutional layers, enabling the refinement network to preserve high-level semantic similarity, facial structure, and visual consistency between generated and reference avatars during high-resolution synthesis.

$$\mathcal{L}_{perc} = \sum_k \|\phi_k(x) - \phi_k(\hat{x}_{128})\|_2 \quad (12)$$

Here, **x** refers to the pixel intensity of the stage 2 generated avatar. To reduce noise artifacts and encourage spatially smooth outputs, total variation regularization is calculated using (13):

$$\mathcal{L}_{TV} = \sum_{i,j} \left(|x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j}| \right) \quad (13)$$

The complete stage 2 optimization objective is achieved using (14),

$$\mathcal{L}_{Stage2} = \mathcal{L}_{adv}^{(2)} + \lambda_{cls}\mathcal{L}_{cls}^{(2)} + \alpha\mathcal{L}_{L1} + \beta\mathcal{L}_{FM} + \gamma\mathcal{L}_{perc} + \delta\mathcal{L}_{TV} \quad (14)$$

λ_{cls}, α, β, γ, and δ denote weighting coefficients that control the relative contribution of each loss term during training. This hybrid loss computations balance visual realism, structural fidelity, and semantic emotion alignment, while improving training stability and suppressing high-frequency artifacts. As a result, stage 2 produces visually refined, perceptually consistent, and emotion-aligned high-resolution avatar representations suitable for immersive AR and VR applications.

III. EXPERIMENTAL RESULTS

This section presents the stage-wise experimental evaluation of the proposed PersonaGAN framework in terms of both qualitative and quantitative results.

A. Refined Conditional Avatar Generation and Evaluation

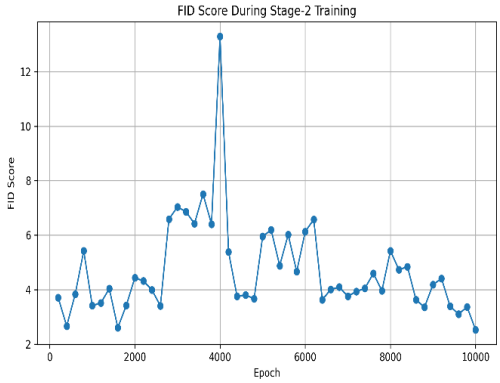
Stage 2 improves the sharpness, color consistency, and structural coherence to create visually refined avatars. Unlike stage 1 output's representation 64 × 64, stage 2 refines these representations into 128 × 128 avatars with enhanced visual fidelity. Table 3 presents the outputs of stage 2 based on four emotion classes (i.e, amusement, baseline,

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

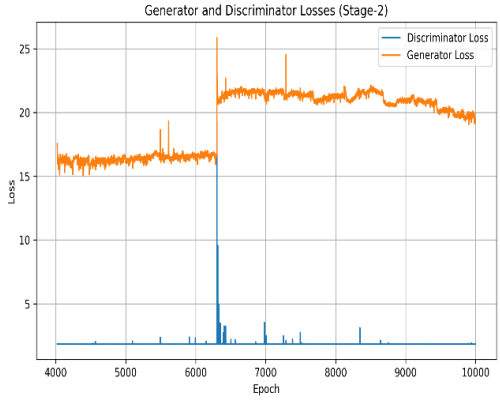
Research Article

meditation, and stress). For each class, two avatars are generated using distinct latent noise samples while conditioning on physiological vectors. This demonstrates the stable emotion-aware avatars generation with controlled intra-class variation. Here cond_idx refers to the dataset index of the conditioning vector containing the corresponding smartwatch physiological features and emotion conditioning information used during avatar generation. Thus, the refinement stage significantly reduces noise, but minor texture artifacts may persist in some cases.

As shown in Fig. 3(a), the training stability and perceptual quality are assessed using FID and adversarial loss curves. Where the FID exhibits the characteristic oscillatory behavior commonly observed in adversarial training, while progressively reaching minimum values close to 2 on the constructed avatar dataset, it indicates a high perceptual similarity between the generated and real samples. Similarly, the generator and discriminator loss curves in Fig. 3(b) present the stable adversarial convergence throughout the training process. Besides, conditional consistency was quantitatively evaluated using an in dependent ResNet18-based emotion recognition model trained exclusively on real avatar images using separate training, validation, and held-out test splits. The classifier demonstrated strong performance on the constructed avatar dataset, achieving high precision and class separability across the predefined emotion categories. The elevated classification accuracy is primarily attributed to the controlled nature of the generated avatar dataset, where emotion categories exhibit visually distinguishable appearance patterns and reduced inter-class ambiguity. After training, the frozen classifier was applied to PersonaGAN-generated avatars to evaluate conditional semantic consistency. The resulting confusion matrix in Fig. 4 demonstrates strong inter-class separability among the generated emotion-aware avatars, with only minor misclassification observed for a small number of stress samples toward the amusement and baseline categories. ResNet18 was an evaluation tool. It was not part of PersonaGAN training and generated avatars were never used to train it.



(a) FID progression



(b) Generator and discriminator losses.

Fig. 3 Stage 2 Training Dynamics.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

TABLE III STAGE 2 LATENT VARIATION UNDER FIXED PHYSIOLOGICAL CONDITIONING

Emotion	Avatar 1	Avatar 2	cond_idx
Amusement			784342, 617417
Baseline			1889154, 2793237
Meditation			3862476, 4639842
Stress			5360465, 5224635

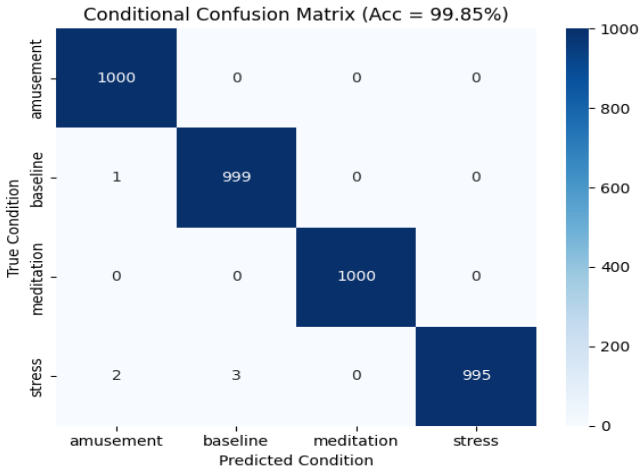


Fig. 4 Conditional confusion matrix using ResNet18 trained on real avatars.

B. Ablation Study on Conditioning Strategies

Ablation studies were performed to understand how physiological features and affective labels influence PersonaGAN. Tested how three different conditioning strategies: labels-only, physiology-only, and combined, impact the model. All the experiments were done with the same architecture and training settings in Stage I for 1000 epochs, and we used the Fréchet Inception Distance (FID) to measure performance. Running with only labels, the model got an FID of 58.93. It thus generated emotion-consistent avatars but they had quite a bit of noise and less visual quality. With physiology only, the model produced really noisily-looking images and got the lowest performance with an FID of 261.78. This means physiological features alone do not give enough semantic guidance to generate avatars reliably. When combining the conditioning strategies, the model scored the best with an FID of 13.54 at 900 epoch. It was able to generate avatars that were both visually better and emotionally consistent. What can be inferred from these findings is that emotion labels offer the main semantic guidance. Physiological features, on the other hand, provide additional information that helps to improve image quality and diversity. The fact that the combined conditioning strategy outperformed the others highlights the value of integrating physiological and affective cues for emotion-aware avatar generation.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

IV. CONCLUSION

This paper introduces PersonaGAN, a health-aware conditional stacked generative adversarial framework that translates physiological signals obtained from a smartwatch into emotion-aware visual avatars. While traditional wearable emotion recognition systems mainly give numerical labels or stress scores, the novel framework allows users to see the internal emotional states visually through avatar synthesis. The variable physiological features and the emotional labels are combined into a single conditioning representation in PersonaGAN, while a two-stage stacked GAN architecture is utilized to produce avatars that are both semantically referring and visually appealing. Through experimental validation, it was found that the adversarial training was stable, the conditional consistency was high across various emotion categories, and the perceptual quality was very good. The ablation research indicated that using physiological features together with affective labels leads to better performance than using only labels or only physiology as conditioning strategies. This also emphasizes the role of physiological information as a complement in emotion-aware avatar creation. Such abilities could provide ways for mental health support systems, telehealth communication, rehabilitation systems, stress-awareness platforms, affective human–computer interaction, and emotion-aware digital assistants, where visual feedback might be more accessible and understandable than numerical indicators. Research will next focus on synthesizing humanized avatars, natural language processing-enabled conversational avatars, and AR/VR integration.

ACKNOWLGEDMENT

The work presented in this paper is carried out as a part of the sponsored research project with grant number No. REV-DM-F008/2/2023-REV-DM-R&D-Part-II, Sr. No. 5 funded by Himachal Pradesh State Disaster Management Authority (HPSDMA) under the State Disaster Mitigation Fund (SDMF).

REFERENCES

[1] J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014)..

[2] M.Mirza, S. Osindero, Conditional generative adversarial nets (2014). arXiv:1411.1784. URLhttps://arxiv.org/abs/1411.1784.

[3] Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. N. Metaxas, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5907–5915.

[4] Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. N. Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks, IEEE trans actions on pattern analysis and machine intelligence 41 (8) (2018) 1947–1962.

[5] Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, Generative adversarial networks: An overview, IEEE signal processing magazine 35 (1) (2018) 53–65.

[6] P. Purwono, A. N. E. Wulandari, A. Ma'arif, W. A. Salah, Understanding generative adversarial networks (gans): A review, Control Systems and Optimization Letters 3 (1) (2025) 36–45.

[7] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, K. Van Laerhoven, Introducing wesad, a multimodal dataset for wearable stress and affect detection, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 400–408. doi:10.1145/3242969.3242985. URLhttps://doi.org/10.1145/3242969.3242985

[8] S. Benita, A. S. Ebenezer, L. Susmitha, M. Subathra, S. J. Priya, Stress detection using cnn on the wesad dataset, in: 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), IEEE, 2024, pp. 308–313.

[9] Nandini, J. Yadav, V. Singh, V. Mohan, S. Agarwal, An ensemble deep learning framework for emotion recognition through wearable devices multi-modal physiological signals, Scientific Reports 15 (1) (2025) 17263.

[10] H.-S. Choi, Emotion recognition using a siamese model and a late fusion-based multimodal method in the wesad dataset with hardware accelerators, Electronics 14 (4) (2025) 723.

[11] S. Kalateh, L. A. Estrada-Jimenez, S. Nikghadam-Hojjati, J. Barata, A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges, IEEE Access 12 (2024) 103976–104019. doi:10.1109/ACCESS.2024.3430850.

[12] Y. Chen, C. Ye, Z. Jiang, J. Li, Y. Wang, T. Ma, N. Xu, Emovr: Guiding and visualizing emotions in virtual reality therapy for mental health, in: 2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), IEEE, 2024, pp. 281–287.

[13] M. Mihelj, D. Novak, M. Munih, Emotion-aware system for upper extremity rehabilitation, in: 2009 Virtual Rehabilitation International Conference, IEEE, 2009, pp. 160 165.

[14] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: A review, Medical image analysis 58 (2019) 101552.

[15] K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, C. S. Greene, Privacy-preserving generative deep neural networks support clinical data sharing, Circulation: Cardiovascular Quality and Out comes 12 (7) (2019) e005122.

[16] S. Tripathi, A. I. Augustin, A. Dunlop, R. Sukumaran, S. Dheer, A. Zavalny, O. Haslam, T. Austin, J. Donchez, P. K. Tripathi, et al., Recent advances and application of generative adversarial networks in drug discovery, development, and targeting, Artificial Intelligence in the life Sciences 2 (2022) 100045.

[17] S. Lee, S. Kim, Y. Chu, J. Choi, E. Park, S. S. Woo, Eaegan: Emotion-aware emoji generative adversarial network for computational modeling diverse and fine-grained hu man emotions, IEEE Transactions on Computational Social Systems 11 (3) (2023) 3862–3872.

[18] Z. Zhang, Y. Liu, S.-h. Zhong, Ganser: A self-supervised data augmentation framework for eeg-based emotion recognition, IEEE Transactions on Affective Computing 14 (3) (2022) 2048–2063.

[19] Polyak, Y. Taigman, L. Wolf, Unsupervised generation of free-form and parameterized avatars, IEEE transactions on pattern analysis and machine intelligence 42 (2) (2018) 444–459.

[20] S. Xu, L. Li, L. Shen, Y. Men, Z. Lian, Your3demoji: Creating personalized emojis via one-shot 3d-aware cartoon avatar synthesis, in: SIGGRAPH Asia 2022 Technical Communications, SA '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 3955–3964. doi:10.1145/3550340.3564220. URLhttps://doi.org/10.1145/3550340.3564220

[21] J. H. Koh, S. Y. Tan, M. F. Nasrudin, A systematic literature review of generative adversarial networks (gans) in 3d avatar reconstruction from 2d images, Multimedia Tools and Applications 83 (26) (2024) 68813–68853. 14

[22] Z. Li, L. Chen, C. Liu, F. Zhang, Z. Li, Y. Gao, Y. Ha, C. Xu, S. Quan, Y. Xu, Animated 3d hu man avatars from a single image with gan-based texture inference, Computers & Graphics 95 (2021) 81–91. doi:https://doi.org/10.1016/j.cag.2021.01.002.

[23] T. Gao, B. Yin, H. Feng, Z. Huang, Instanthuman: Single-image to high-fidelity 3d human in under one second, Computers & Graphics 133 (2025) 104464. doi:https://doi.org/10.1016/j.cag.2025.104464.

[24] Liu, X. Wang, Z. Wan, Y. Ma, J. Chen, Y. Fan, Y. Shen, Y. Song, Q. Chen, Avatarartist: Open-domain 4d avatarization, in: Proceedings of the Computer Vision and Pat tern Recognition Conference, 2025, pp. 10758–10769.

[25] M. Chen, D. Liu, S. Ma, M. Vasilkovsky, B. Zhou, Q. Gao, W. Wang, J. Luo, D. N. Metaxas, V. Sitzmann, et al., Snapmoji: Instant generation of animatable dual stylized avatars, arXiv preprint arXiv:2503.11978 (2025).

[26] Y. Ding, J. Liu, W. Zhang, Z. Wang, W. Hu, L. Cui, M. Lao, Y. Shao, H. Liu, X. Li, et al., Kling avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis, arXiv preprint arXiv:2509.09595 (2025).

[27] R. Wang, Y. Cao, K. Han, K.-Y. K. Wong, A survey on 3d human avatar modeling–from reconstruction to generation, arXiv preprint arXiv:2406.04253 (2024).

[28] Y. Miao, Y. Liu, Advances in vision-based deep learning methods for interacting hands reconstruction: A survey, Computers & Graphics 124 (2024) 104102. doi:https://doi.org/10.1016/j.cag.2024.104102.

[29] H. He, G. Li, Z. Ye, A. Mao, C. Xian, Y. Nie, Data-driven 3d human head reconstruction, Computers & Graphics 80 (2019) 85–96. doi:https://doi.org/10.1016/j.cag.2019.03.008.

[30] W. Woodworth, C. W. Borst, Visual cues in vr for guiding attention vs. restoring attention after a short distraction, Computers & Graphics 118 (2024) 194–209. doi:https://doi.org/10.1016/j.cag.2023.12.008.

[31] DiceBear Team, Dicebear avatars api, https://dicebear.com/api, accessed: Jan 2026 (2025).

[32] A.Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 2642–2651.