# Generative AI for Human 3D Body Emotions: A Dataset and Baseline Methods

Ciprian Paduraru[1], Petru-Liviu Bouruc[1] and Alin Stefanescu[1,2]
[1] *Department of Computer Science, University of Bucharest, Academiei Street 14, Bucharest, Romania*
[2] *Institute for Logic and Data Science, Romania*

Keywords: Generative AI, Body Emotions, Parametric Models, Animations.

Abstract: Accurate and expressive representation of human emotions in 3D models remains a major challenge in various industries, including gaming, film, healthcare, virtual reality and robotics. This work aims to address this challenge by utilizing a new dataset and a set of baseline methods within an open-source framework developed to improve realism and emotional expressiveness in human 3D representations. At the center of this work is the use of a novel and diverse dataset consisting of short video clips showing people mimicking specific emotions: anger, happiness, surprise, disgust, sadness, and fear. The dataset was further processed using state-of-the-art parametric body models that accurately reproduce these emotions. The resulting 3D meshes were then integrated into a generative pose generation model capable of producing similar emotions.

## 1 INTRODUCTION

This work addresses the challenge of enhancing the emotional expressiveness of 3D human models by proposing a novel framework. The goal is to generate 3D representations of human bodies that accurately mimic the emotional expression of real life. To achieve this, we first needed to identify a dataset that is expressive enough to allow the extraction of 3D meshes in different poses corresponding to different emotions.

**Dataset.** While there are several datasets available online, they either do not capture the whole body, as in datasets such as (Sun et al., 2021), (Mollahosseini et al., 2017) and (Zadeh et al., 2018a), or they focus on actions rather than emotions, such as clapping in the (Zadeh et al., 2018b) dataset.

- We have developed a unique dataset that captures the full body posture of people expressing six specific emotions in real time, including the transitions between a normal (relaxed) posture and each of these postures. This dataset overcomes the limitations of existing datasets by capturing the full expression cycle and allowing for natural emotional expressions.
- Adapted several of the state-of-the-art methods that process poses and meshes in both 2D and 3D space and assembled them into a pipeline process.
- Proposed a new generative process that starts from any body mesh (even outside the dataset) of a person and generates one of the six emotions our models are trained on: anger, happiness, surprise, disgust, sadness and fear.
- The methods, dataset, and experiments are published as open source on https://github.com/unibuc-cs/3DHumanEmotionsGenerator for further research in academia and industry.

## 2 RELATED WORK

The representation of human emotions in the digital realm has advanced significantly in recent decades. Early efforts, such as those in (Vrajitoru, 2006), explored digital chatterbots and NPCs capable of human-like interactions. With the advent of technologies like deep learning and generative AI, research has expanded, as shown in (Li et al., 2023) and (Park et al., 2023), which trained AI models inspired by neuroscience and simulated interactive bot environments. Current research emphasizes large language models (LLMs), which excel at complex tasks and conversational agent simulations.

Efforts to replicate human 3D representations have also evolved. Early works (Pons-Moll et al., 2015), (Kolotouros et al., 2019), (Anguelov et al.,

Paduraru, C., Bouruc, P.-L. and Stefanescu, A.
Generative AI for Human 3D Body Emotions: A Dataset and Baseline Methods.
DOI: 10.5220/0013168700003890
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 646-653
ISBN: 978-989-758-737-5; ISSN: 2184-433X

2005) laid the foundation, with SCAPE (Anguelov et al., 2005) offering a detailed model for representing human bodies in 3D. SCAPE combines shape- and posture-dependent deformations, utilizing principal component analysis (PCA) for efficiency. Extensions like Blend-SCAPE (Hirshberg et al., 2012) introduced smoother transitions, paving the way for models like SMPL (Loper et al., 2015) and SMPLify (Bogo et al., 2016), which simplify and improve 3D human modeling.

Alternative approaches, such as diffusion-based models, emerged in parallel. ScoreHMR (Stathopoulos et al., 2024) leverages diffusion models and score-based learning to overcome optimization challenges, enhancing 3D mesh recovery. It refines noisy estimates of 3D meshes through iterative denoising, incorporating multi-view refinement and motion data for dynamic scenes, achieving high accuracy.

Pose estimation, a critical challenge in 3D human modeling, is addressed by OpenPose (Cao et al., 2019), (Zhao et al., 2024). Its two-stage architecture combines CNN-generated confidence maps and part affinity fields (PAFs) to connect keypoints into skeletons, supporting multi-person scenarios and hand and facial point recognition.

The novelty of this research lies in integrating and adapting these techniques to develop a comprehensive system for realistic whole-body human emotion representation using a new dataset, addressing gaps not tackled by existing models.

# 3 THEORETICAL FOUNDATIONS

## 3.1 Representation of the Human Body

The SMPL (Loper et al., 2015) model represents the human body as a 3D mesh consisting of about 6,890 vertices, which corresponds to about 20,670 floats describing the entire body. The goal of SMPL is to define a function $M(\bar{\beta}, \bar{\theta})$, where $\beta$ stands for the shape parameters and $\theta$ for the pose parameters. The function learns to map the input parameters to human 3D body meshes (as mentioned above) to ensure that the output represents valid, realistic body configurations. This ability stems from the model's ability to capture variations in shape and pose, which enables the parameters. And this is the final formula for the SMPL model:

$$M(\bar{\theta}, \bar{\beta}) = W\left( \underbrace{T_P(\bar{\beta}, \bar{\theta})}_{\text{Deformed template mesh}}, \underbrace{J(\bar{\theta})}_{\text{Joints}}, \underbrace{\mathcal{W}}_{\text{Skinning weights}}, \bar{\theta} \right) \quad (1)$$

where

$$T_P(\bar{\beta}, \bar{\theta}) = \bar{T} + \underbrace{B_S(\bar{\beta})}_{\text{Shape deformation}} + \underbrace{B_P(\bar{\theta})}_{\text{Pose deformation}}$$

- $\bar{T}$ - the average template mesh;
- $J(\bar{\theta})$ - the 3D positions of the skeletal joints that control posture, based on the pose parameters $\bar{\theta}$;
- $\mathcal{W}$ - parameters learned from data (training was performed on 1786 3D scans of humans in different poses)
- $W$ - Linear Blend Skinning (LBS) function used to deform the mesh (the 3D human body) based on joint rotations and skinning weights.

To improve the usability of the SMPL model, an important task is to extract 3D meshes from images without additional inputs such as camera parameters or pose data. The method used in our work is based on **SMPLify** (Bogo et al., 2016), which combines 2D joint detections (2D pose) obtained with methods such as OpenPose (Cao et al., 2019) with the output of SMPL. The goal is to match the 3D joints generated by the SMPL model with the 2D pose estimates by minimizing an objective function. One of the challenges is self-penetration, where body parts overlap. SMPLify addresses this problem by introducing a penalty for self-intersection that prevents unrealistic overlaps. Another challenge, depth ambiguity, is solved using pose priors from the SMPL training data that penalize implausible poses to improve the accuracy of the estimation. An example from our application is shown in Figure 2. Further details on the implementation can be found in (Bogo et al., 2016).

The next tool used is **SMPL-X** (Pavlakos et al., 2019), which not only models the entire human body in 3D, but also facial expressions and hand movements. These two areas are crucial for human communication, and SMPL-X increases realism by including detailed expressions for both. To capture the details of the face and hand, the authors used specialized datasets and models:

- a model to represent head meshes. It uses FLAME (Bolkart and Wuhrer, 2021), which is based on 3,800 head scans.

- a model to represent the hand. The model MANO (Romero et al., 2017) is used, which is based on 1,500 hand scans.

Similar to SMPL, principal component analysis (PCA) was applied to the above datasets to extract the principal components. Additionally, the number of joints in the model was increased from 23 to 54 to account for the added complexity of the head and hand networks. SMPL-X distinguishes between male, female and, where necessary, gender-neutral body shapes.

SMPLify-X (Pavlakos et al., 2019) introduces several improvements, including VPoser (a variational autoencoder that learns a distribution over likely human poses), a refined interpenetration penalty, an improved gender detector, and an overall faster implementation.

### 3.2 Generative Model

The purpose of this model is to move values in the variety of open inputs to generate diverse and high quality postures. In this sense, VPoser (Pavlakos et al., 2019) compresses body poses into a low-dimensional latent space and reconstructs valid poses from this space, ensuring alignment between the joints of the 3D model and the 2D joints recognized by OpenPose. This is achieved by minimizing an objective function that relates the two sets of joints and optimizes the pose accuracy of the model.

VPoser is a deep learning-based body pose prior designed to model and regularize human 3D poses for various applications such as animation, virtual reality, robotics and medicine. Based on a variational autoencoder (VAE), VPoser learns a latent space representation of human postures that enables both pose synthesis and probabilistic inference. The latent space encodes the most important features of body movements and postures and ensures that the generated or reconstructed poses are realistic and correspond to the natural movement constraints of humans. One of the main advantages of VPoser is its ability to provide smooth and consistent prioritization for pose generation, making it particularly effective at minimizing physically implausible poses.

VPoser is trained using an extensive dataset of human poses that includes various sources of motion capture data to ensure diversity and coverage of a wide range of human movements. In particular, the training data includes motion capture poses from publicly available datasets, including the CMU motion capture database Human3.6M (Ionescu et al., 2014) and the PosePrior dataset (Akhter and Black, 2015). These datasets are processed using MoSh (Motion and Shape capture ) (Loper et al., 2014), a technique that extracts pose parameters in SMPL model format to ensure compatibility with 3D body shape models. The resulting pose parameters are then represented in the form of rotation matrices, which are commonly used in computer graphics and machine learning for accurate and smooth rotational transformations. The VPoser model was trained on approximately one million pose samples, with a separate test set of 65,000 poses used to evaluate generalization performance. These poses are represented in the form of rotation matrices to ensure consistency with the SMPL body model.

## 4 METHODS

This section first describes the process of creating the dataset. It then presents the pipeline used to extract each frame of the movie, understand it from a skeletal perspective, and apply body meshes to capture each person. Finally, details are presented on the generative model that can be used to create a full 3D body expression starting from any given human mesh object.

### 4.1 Data Collection

A novel dataset was collected with the help of students and professionals at University of Bucharest. Each participant had to record the expression of different emotions in a place of their choice. The dataset consists of short videos (about 8 seconds each) in which the participants express one of six primary emotions: anger, happiness, surprise, disgust, sadness, and fear. Each video captures the progression of the emotion, starting from a neutral state, through the full expression of the emotion, to the return to neutrality, providing a comprehensive representation of the emotional dynamics. For each person and each emotion, 10 videos were recorded showing the entire body. To ensure optimal mesh extraction for generating human-like behaviors for agents, the videos had to meet several technical requirements. These included a minimum resolution of FullHD and a simple, preferably monochrome background (e.g. white). These conditions were set to avoid complications such as blurred images or unclear contours that could affect the accuracy of the shape extraction.

A key advantage of this dataset is that it focuses on capturing a wide range of emotional expressions directly from real-life videos. By allowing people to express their emotions in their own unique way, the dataset provides a rich diversity of emotional behav-

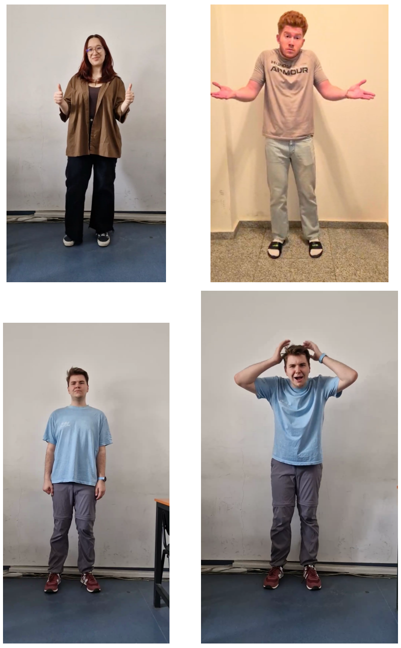

Figure 1: Snapshots of people and reactions from our data set. In order of emotions shown such as happy, surprised, neutral and scared.

ior, which results in the embedding later being highly expressive and representative of real human reactions. This diversity is essential for creating realistic 3D poses and helps ensure that the dataset captures a wide range of human emotions, which is crucial for applications in games, animations and virtual reality environments.

By extracting SMPL-X objects from video frames, the dataset provides a variety of poses that accurately represent the corresponding emotional expression. This variety is critical for future applications, especially in production environments such as the Unreal Engine, where the ability to recreate realistic, human-like behavior for 3D characters is critical to simulation realism and user perception. By linking different poses - from neutral expressions to emotional state to return to neutrality - this dataset facilitates the development of dynamic, believable character animations that can mimic real-life emotional responses, significantly improving immersion and engagement.

## 4.2 Extracting Meshes

Once the samples have been collected, the first goal is to extract the 3D meshes representing the different poses. We chose SMPLify-X (Pavlakos et al., 2019) as the primary framework because of its ability to capture body, face and hand expressions. More specifically, the implementation of SMPLpix and SMPLify-X (Prokudin et al., 2021) was reused and adapted it for our dataset. A concrete representation of this pipeline can be found in Figure 2. In this pipeline, after experimentation, we created a middle step to recognize the skeleton from the image frame using the OpenPose (Cao et al., 2019) (Simon et al., 2017) (Cao et al., 2017) (Wei et al., 2016) solution, as it was able to capture not only the skeleton but also the intricate details of the hands and face that we further required for our goals. In addition to the image sequence from the video, the input for this pipeline optionally includes the gender specification. Once the process is complete, three key files are output:

- an augmented image showing the mesh and skeleton.
- the 3D mesh structure, an .obj file.
- additional mesh data for further analysis and processing, a file archived in a disk file.

## 4.3 Generation of New Poses

When creating realistic, human-like 3D mesh representations, one of the final steps is to generate poses and then condition them on a specific body and adapt them to certain categories and parameters, Figure 3. To achieve this goal, a Variational Autoencoder (VAE) based method proved to be the most suitable approach from the evaluation.

Initially, the experiments started with training a VAE from scratch. However, the lack of sufficient training data was a major obstacle to generating high quality results. Extracting a single 3D mesh is very computationally intensive, making it impractical to generate a large enough dataset to train the model effectively. Given these limitations, we experimented with transfer learning of VPoser (Pavlakos et al., 2019), a VAE specifically designed for the SMPL model and already trained on a large amount of data. We reused the basic ideas of the VPoser method to input pre-generated 3D meshes into the VAE and create similar mesh variants. The original mesh was encoded into the latent space of the VAE, where transformations were applied to the latent vectors. In particular, small perturbations were added. The modified latent vectors were then decoded to generate new similar meshes. By controlling these transformations, variations of the original mesh were generated while preserving its main structural properties. However, we adapted the original implementation to represent a Gaussian model for the difference between two consecutive poses based on the frame and the

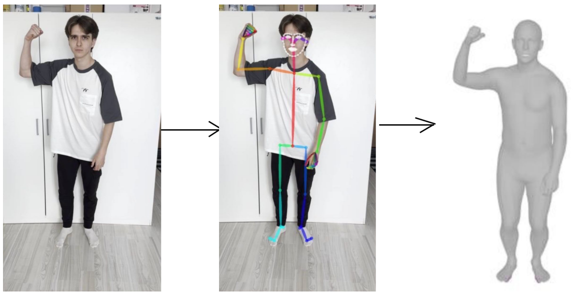

Figure 2: Example for the application of the pipeline to the data set: OpenPose - SMPLify-X. From left to right, OpenPose is first applied to each image extracted from each video in the dataset. The resulting image in the center contains the person with the 3D skeleton on top. With SMPLify-X, we get the full 3D body skeleton in the right part.

previous pose. This model was trained with the proposed dataset. A pseudocode of our modified version of this process is shown in Listing 1. The function $getRandomPoseDiff(t)$ extracts from the Gaussian model mentioned above, constrained by the previous frame and the time of the sequence.

Listing 1: Sample Python Code.

```
def generate_pose(initial_pose ,
                  num_frames):
 # the noise scale
 ns = 0.5
 Pose_{0} = initial_pose
 for t in 1..num_frames:
    $P_a$ =getRandomPoseDiff(t, $Pose_{t-1}$)
    $Pose_t$ = $Pose_{t-1}$ + ns * $P_a$
```

In Listing 1, the *noise_scale* represents the degree of variation introduced into the system and controls the extent to which the output deviates from the original. In particular, it indicates the extent of the change after the decoding process. In combination with the original posture parameters, this variation results in the generation of a modified version of the mesh represented by the associated parameters. This process allows the controlled exploration of different mesh configurations while maintaining a relationship to the original pose.

## 5 EVALUATION

In this section, we evaluate our work from different perspectives, to which we have mainly contributed. We start with the obtained dataset, evaluate the pipeline with the mesh extraction and then the generative model. Finally, we present an ablation study from our experiments with other models, discuss the limitations, observed artifacts and the general applicability of our work. The computational resources required for training and inference are also discussed.

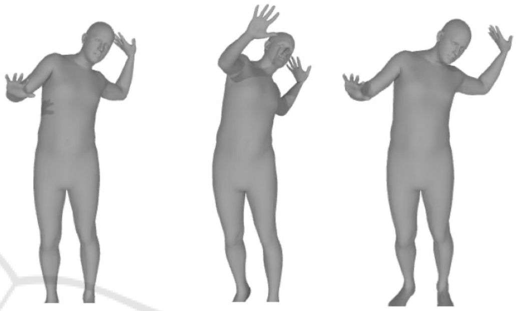

Figure 3: A generative process in action in our framework. Left: the initial mesh; center: the intermediate result mesh. Right: the final generated pose.

### 5.1 Dataset: Metrics and Challenges

The Table 1 contains general data about how the dataset is organized and structured.

Table 1: Dataset overview.

| | value |
|---|---|
| participants | 28 |
| female participants | 6 |
| male participants | 22 |
| expressions | 6 |
| videos per expression | 9 - 10 |
| total number of videos | 1677 |
| seconds per video | 6 - 10 |

One of the biggest challenges in creating a new dataset is the manual work that needs to be done after the initial recordings. In our case, people have mimicked emotions that look more like different labels than the ones we were originally looking for, e.g. *anxiety*. We decided to isolate these to ensure that we only considered good quality data. In addition, there were subjective interpretations of emotion expressions, such as people expressing *fear* instead of *scared*, which raised questions about the subtle differences between these two emotions (fear is an enduring emotion, while anxiety is often a more immediate reaction). Some videos also had technical limitations, such as poor quality or problematic viewing angles that could make it difficult to extract meshes. Despite these issues, most of the submitted videos were retained and provide a solid basis for creating expressive 3D models from real expressions of emotion.

## 5.2 Evaluation of the Mesh Extraction

Our experiments have shown that the pipeline discussed in Section 4 accurately captures the overall body shape as well as detailed poses for the body, face and hands (with the hand being particularly well captured, as shown in 2).

Table 2: Quantitative comparison of the average performance of SMPLify-X on different subsets of the dataset and on the whole dataset. V2V (vertex-to-vertex) error refers to the error metric used to evaluate how well the estimated human 3D shape matches the ground truth.

| Dataset | avg v2v error |
|---|---|
| anger samples | 53.2 |
| happiness samples | 54.1 |
| surprise samples | 51.4 |
| disgust samples | 53.4 |
| sadness samples | 52.3 |
| fear samples | 50.9 |
| female models | 52.9 |
| male models | 52.5 |
| Overall | 52.6 |

The results obtained above in 2 confirm that the proposed dataset and framework have successfully achieved their goal of extracting expressive 3D meshes from custom videos. These extracted meshes can subsequently be used to generate similar poses and used for practical applications such as NPC training or virtual human representations in real-world scenarios.

## 5.3 Generative Model Evaluation

Based on our experiments, the generative model demonstrated the ability to generate high-quality outputs with valid and realistic poses. When evaluated using the vertex-to-vertex (V2V) error metric, the generated poses showed a strong correspondence to the original input meshes, indicating minimal deviation in vertex positions. This quantitative assessment confirms that the outputs produced by generative model are not only plausible in terms of pose but also sufficiently close to the original meshes, validating the model's performance in generating accurate and reliable results.

Figure 4 illustrates the relationship between vertex-to-vertex similarity (V2V) and increasing noise applied to the generated mesh. As the noise level that perturbs the vertex positions increases, the V2V similarity metric decreases significantly. This is because higher noise leads to larger deviations from the original mesh, which in turn leads to larger Euclidean distances between the corresponding vertices of the two meshes. Consequently, as the noise increases, the generated mesh deviates more and more from the original, leading to a progressive reduction in overall similarity. The spikes (either up or down) are caused by the randomness that occurs when new deviations are generated. In this way, it can be shown that the generative model was able to produce meaningful poses, which could then be linked to specific emotional categories or other criteria such as posture or movement.

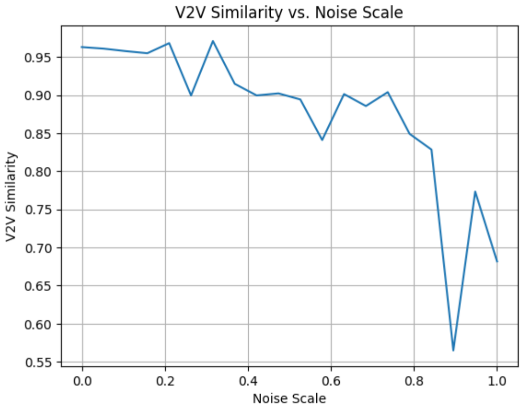

Figure 4: V2V similarity decreases as noise levels increase, indicating greater divergence between the generated and original meshes due to larger deviations in vertex positions.

## 5.4 Artifacts

Overall, the extracted meshes are even if the images are of good quality, there are some poses that can produce unwanted artifacts.

In Figure 5, where OpenPose successfully extracted the hand position in front of the face, we observe a discrepancy in the application of the framework: the resulting mesh seems to place the hand behind the head. To our understanding, this kind of misalignment is probably due to the function responsible for aligning the pose generated by the Variational Autoencoder (VPoser) with the pose extracted by OpenPose. The misalignment could potentially be fixed by retraining the entire model on the dataset used to ensure better synchronization between the generated and extracted poses. However, retraining the components of the pipeline from scratch instead of doing transfer learning would require a significant amount

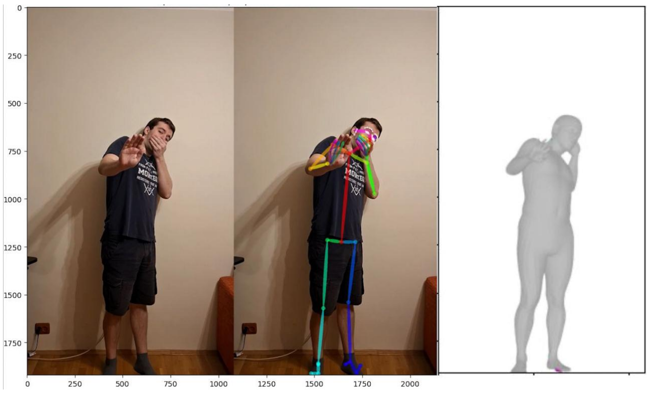

Figure 5: Artifact in the pipeline: the mesh (right) seems to place the hand behind the head, in contrast to the real image on the left, where the hands are in front of the head. The skeleton detection phase (center) is correct.

of computational resources.

## 5.5 Comparison with Other Models

For comparison, we also explored mesh extraction with ScoreHMR by implementing the code provided by the authors in (Stathopoulos et al., 2024). In our initial observations, we noticed a significant improvement in the speed of mesh extraction from images compared to SMPLify-X. In particular, in tests with demo images containing multiple subjects, ScoreHMR processed the images much faster (about 5x faster for an image with multiple subjects). This increase in speed is due to the use of diffusion models that analyze the entire image at once, as opposed to classical optimization methods that extract meshes sequentially. However, a notable drawback of ScoreHMR is that it sometimes fails to generate the correct pose (as in Figure 6). Also, it currently only supports the SMPL model, meaning it cannot extract face or hand poses, which limits its overall versatility.

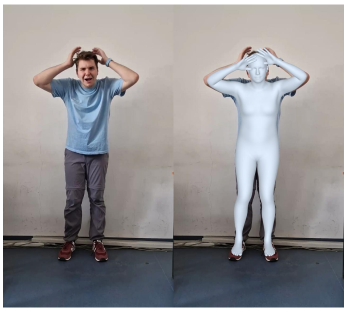

Figure 6: 3D mesh, extracted with ScoreHMR. Although the mesh was extracted in less time compared to SMPLify-X, it cannot reproduce the facial and hand expressions and still outputs the position of the hands incorrectly.

## 5.6 Discussion of Applicability

In our experiments, the computational performance of the proposed pipeline and our generative model was evaluated to estimate the time required for mesh extraction and pose generation. With an A100 GPU, the pipeline processes a single frame in about 52 seconds. For a short video of 10 seconds duration and 30 frames per second, the extraction of the corresponding meshes would therefore take about 4 hours and 20 minutes. Given the size of our dataset, the total time to process all frames is estimated to be about 7267 GPU hours (or 302 days and 19 hours for a single GPU), based on 1. However, the processing scales almost optimally with the number of GPUs used. Furthermore, the generative model generates meshes at a rate of 2 seconds, which means that creating $N$ frames of an animation would take approximately $Nx2$ seconds. A concrete example: Creating a new animation of 10 seconds at a rate of 30 frames per second requires $\sim$ 10 minutes.

Offline creation of emotion animations is important for the industry as it eliminates real-time overhead and is equivalent to manually created animations. For example, a game developer can use the proposed model to create and save animations, which significantly reduces costs and simplifies asset management and editing. In addition, the full 3D mesh on each frame provides the necessary detail for the integration of skin, shaders and clothing, resulting in fully renderable characters.

# 6 CONCLUSIONS

This research deals with the creation of 3D animations that represent human emotions, which are essential in areas such as games, film, healthcare, virtual reality and robotics. An open-source dataset of various emotional expressions was developed and a processing pipeline was implemented to analyze skeletal and 3D body representations. A generative model based on Variational Autoencoders (VAEs), in particular VPoser, was used to generate new 3D poses that retain emotional nuances. Future work includes integrating these poses into NPC animation pipelines, extending the dataset for better visualization, and evaluating the impact on user experience in real-world applications.

## ACKNOWLEDGEMENTS

This research is partially supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906 and a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI-UEFISCDI, project no. PN-IV-P8-8.1-PRE-HE-ORG-2023-0081, within PNCDI IV.

## REFERENCES

Akhter, I. and Black, M. J. (2015). Pose-conditioned joint angle limits for 3D human pose reconstruction. In *In proceedings of IEEE CVPR 2015*.

Anguelov, D. et al. (2005). Scape: shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416.

Bogo, F. et al. (2016). Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing.

Bolkart, T. and Wuhrer, S. (2021). FLAME: A 3d morphable model of the head and face based on 3d scans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):2808–2821.

Cao, Z. et al. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.

Hirshberg, D. et al. (2012). Coregistration: Simultaneous alignment and modeling of articulated 3d shape. 7577:242–255.

Ionescu, C. et al. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339.

Kolotouros, N. et al. (2019). Learning to reconstruct 3d human pose and shape via model-fitting in the loop.

Li, C. et al. (2023). The good, the bad, and why: Unveiling emotions in generative ai. In *ICML 2024*.

Loper, M. et al. (2015). Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6).

Loper, M., Mahmood, N., and Black, M. J. (2014). Mosh: motion and shape capture from sparse markers. *ACM Trans. Graph.*, 33(6).

Mollahosseini, A., Hassani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *CoRR*, abs/1708.03985.

Park, J. S. et al. (2023). Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Pavlakos, G. et al. (2019). Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Pons-Moll, G. et al. (2015). Dyna: A model of dynamic human shape in motion.

Prokudin, S., Black, M. J., and Romero, J. (2021). Smplpix: Neural avatars from 3d human models. In *Proceedings of the IEEE/CVF WACV*, pages 1810–1819.

Romero, J., Tzionas, D., and Black, M. J. (2017). Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):245:1–245:17.

Simon, T. et al. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.

Stathopoulos, A., Han, L., and Metaxas, D. (2024). Score-guided diffusion for 3d human recovery. In *CVPR*.

Sun, J. J. et al. (2021). Eev: A large-scale dataset for studying evoked expressions from video. *arXiv preprint arXiv:2001.05488*.

Vrajitoru, D. (2006). Npcs and chatterbots with personality and emotional response. pages 142 – 147.

Wei, S.-E. et al. (2016). Convolutional pose machines. In *CVPR*.

Zadeh, A. et al. (2018a). Multimodal sentiment analysis of videos: Facial expressions, text, and audio. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5006–5015. Association for Computational Linguistics.

Zadeh, A. et al. (2018b). Multimodal sentiment analysis of videos: Facial expressions, text, and audio. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5006–5015. Association for Computational Linguistics.

Zhao, W. et al. (2024). Open-pose 3d zero-shot learning: Benchmark and challenges.