



GENERATIVE AI FOR CREATING HYPER- REALISTIC 3D HUMAN MODELS

Narayana Gaddam

Department of Technology and Innovation, City National Bank, USA.

ABSTRACT

Generative artificial intelligence (AI) have by far been instrumental in building hyper realistic 3D human models, ensuring their usage in gaming, virtual reality and digital content creation. In this research, we explore techniques of predicting highly accurate and lifelike 3D human avatar using state of the art neural networks approaches. That is to say, the trend of generating high fidelity 3D representations using 2D image collections has been shown possible with recent works like AG3D and Get3DHuman. LumiGAN introduces relightable face generation to help realistic motion of a rendered person in dynamic lighting environment, and SMPLpix brings neural avatar to enable seamless human motion synthesis. This work integrates methods such as the Skinned Multi Person Linear (SMPL) model and 3D Morphable Models (3DMM) to improve the quality of pose, texture and geometry estimation to produce better quality facial and body reconstructions. The research also applies pixel aligned reconstruction priors, to improve spatial consistency and structural precision; and volumetric convolutional neural networks (CNNs) to enhance the quality for the spatial consistency and structural precision. Experimental results show better performance such as tending to generate photorealistic textures, photorealistic dynamical facial expressions, and human poses with little input data. The research promises to further enhance the experience in the area of augmented and virtual reality immersive

experiences through improvements in these advances. This serves as future work that will deal with challenges to performance optimization in real time and cross-domain adaptability. This work is a contribution to trends in 3D model synthesis from concept to application, that fills the gap between computer vision, machine learning, and visual computing.

Keywords: Generative AI, 3D Human Models, Hyper-Realism, Neural Networks, Computer Vision, Virtual Reality, Augmented Reality, Digital Avatars

Cite this Article: Narayana Gaddam. (2022). Generative AI for Creating Hyper-Realistic 3D Human Models. *International Journal of Information Technology and Management Information Systems (IJITMIS)*, 13(1), 137-149.

https://iaeme.com/MasterAdmin/Journal_uploads/IJITMIS/VOLUME_13_ISSUE_1/IJITMIS_13_01_013.pdf

I. Introduction

Nowadays, the creation of hyper realistic 3D human models is an important challenge in computer vision and digital content creation. In order to augment immersive experiences, increasing number of industries like gaming, film production and virtual reality, urge for accurate and as detailed as possible 3D human representation. This is difficult though, as it is challenging to get photorealism in 3D models by solving the texture mapping problem, pose estimation problem, and geometry refinement problem. Current traditional 3D modeling techniques typically involve a great manual effort and lack the capacity to model naturalistic human characteristics.

Generative artificial intelligence (AI) has been recently advancing in terms of powerful frameworks that can automate the synthesis of hyper real 3D human models. The deep neural networks used in these methods including AG3D [1] and Get3DHuman [2] can generate precise 3D human models from 2D image dataset instead of a direct input of human data. Furthermore, models such as LumiGAN [3] are also capable of performing dynamic relighting, which adds to the realism in different lighting environments. Improved motion synthesis for virtual environments is achieved from the introduction of neural avatars like SMPLpix [4].

However, there are still challenges to overcome to generate in real time, to generate consistently across different poses, and to adapt the model to new environments that have not been seen before. In order to address these gaps, this research seeks to explore hybrid generative techniques composed of volumetric convolutional networks [9] in conjunction with pixel-aligned priors [2] to enhance spatial coherence, realism, and model efficiency.

This can also be considered as the contribution to ongoing increment in 3D model synthesis as regards performance, visual fidelity and adaptability for real-life applications.

II. LITERATURE SURVEY

The synthesis of more and more hyper realistic 3D human models by means of generative AI has led to great progress in digital content creation, virtual reality and gaming. Firstly, early approaches simple tried to apply traditional 3D modelling tools by making extensive use of manual efforts, which was both inconsistent and laborious. Nowadays, deep learning frameworks are used as modern techniques to generate very realistic human models with more precision and less manual intervention. Other usage of neural networks as methods such as AG3D [1] and Get3DHuman[2] are able to generate 3D avatars from 2D image collecitons, making the model more realistic and accurate in geometry. Dynamic relighting for better facial textures were introduced by LumiGAN [3] and better motion synthesis via neural avatars was done by SMPLpix [4]. However, real time performance, scalability and capability to handle unseen data as well as detailed feature representation are still the challenges. Given these advances, this research combines volumetric CNNs [9] with pixel aligned priors [2] to achieve higher spatial coherence, realism and model efficiency.

1. Generative AI for 3D Human Model Synthesis

Due to good performance in 3D model generation, GANs and autoencoder are used. The AG3D [1] is a learning framework which synthesizes 3D human avatars from 2D image datasets. Through using adversarial learning, AG3D maintains photorealistic textures and accuracy of the body's proportions. In a similar fashion, Get3DHuman [2] further improves this by incorporating pixel aligned reconstruction priors with a 3D aware generator to get better quality of reconstructed human models. Two of the key limitations around both of these traditional modeling techniques are solved by the two frameworks: reducing the amount of manual input and improving automation. These methods have the potential of generating promising results, but in practice their performance is usually poor at producing consistent results for dynamic motion or non standard poses. This research tries to combine hybrid techniques so to resolve these issues and enhance both model consistency and realism.

2. Dynamic Relighting for Enhanced Visual Fidelity

The realistic lighting has a significant impact on improving the authenticity of 3D human models. In [3], LumiGAN is introduced as a generative architecture for relightable 3D

face generation. LumiGAN synthesizes high resolution facial textures under different lighting conditions and increases the realism of dynamic scenes. A solution to this major problem in providing visual coherence in scene transitions of virtual environments is provided by this technique. Furthermore, given these relightable properties, generative models are applicable even in film, animation, and gaming. Although it works well, LumiGAN has trouble coping with texture degeneration in the presence of severe illumination changes. This method can be combined with pixel aligned priors [2] to improve structural precision on stable and accurate feature mapping in high contrast condition.

3. Neural Avatars for Motion and Pose Estimation

Neural avatars, that is, 3D human models, are introduced in the motion synthesis pipeline in [4] through SMPLpix. SMPLpix represents consistent, sequence consistent and realistic motion sequences by encoding pose and geometry features in a neural framework. It eliminates the need of the traditional rigging techniques and simplified the animation pipeline. We demonstrate that SMPLpix can be used in gaming and virtual reality applications where contamination free human motion is important. But the motion inconsistency and model distortion on extreme poses are still an open issue. SMPLpix will be properly integrated with current state of the art pose estimation frameworks for precise dynamic interactions and to stabilize the avatar.

4. Volumetric CNNs for Structural Consistency

3D human model reconstruction is a problem that has been addressed with volumetric convolutional neural networks (CNNs) as powerful tools. Direct volumetric CNN regression methods [9] encode 3D spatial data whose better feature alignment and structural accuracy is guaranteed. Unlike them volumetric CNNs permit to give comprehensive depth perception, which is perfect to reconstruct complex facial feature and body geometry. The effectiveness of this approach is that it helps in making AR/VR environments and gaming more realistic. Volumetric models, however, generally require a very large computational resource, so they can not be practically implemented in real time applications. Currently there are ongoing research on optimizing these models through lightweight architectures and GPU acceleration.

5. Hybrid Techniques for Improved Model Adaptability

However, to overcome the above mentioned performance scalability and cross domain adaptability challenges, hybrid techniques based on the hybrid of more than one framework are popular. To achieve such improvement on the consistency of texture generation and spatial mapping, researchers have integrated pixel-aligned priors [2] with volumetric CNNs [9]. Stability of the hybrid approach over different environments is bolstered and texture sharpness

is improved. Furthermore, it is possible to generate better geometry predictions for non standard poses through the combination of generative models such as SMPL [6]. These techniques have the potential of achieving real time performance, and therefore are applicable to gaming, animation, and virtual reality applications. More research is necessary to obtain a proper tradeoff between complexity and performance efficiency.

III. MATERIALS AND METHOD

The proposed methodology to generate hyper real 3D human models uses advanced streamlining of generative AI methods with real time performance optimizations. Instead, we implement a hybrid architecture which utilizes volumetric CNNs [9] as well as pixel aligned reconstruction priors [2] for the benefit of spatial coherence, texture accuracy, and realness of the model. The framework is designed to realize the real time synthesis, which uses the GPU acceleration, the pipeline of data, and the model structures that are optimized.

This implementation is carried out in a high performance computing environment with an AMD Ryzen 7 5800H processor with Radeon Graphics, 16 GB RAM and an NVIDIA GeForce RTX 3060 GPU. The configuration chosen for this project guarantees a sufficient computational power capable of processing volumetric data, train complex generative models, and dealing with dynamic rendering on real-time scenarios. It is set up on Ubuntu 22.04 with CUDA 11.8 for GPU acceleration to run with maximum performance during training and inference. All software frameworks are based on the framework Python, PyTorch, and TensorFlow, doing a good job handling deep learning architectures and working with experimental simulations.

The developed multiresolution encoding layers enable volumetric CNN backbone for model implementation that capture depth information and structural details at higher precision [9]. Aligning 2D feature maps (in different views) to 3D spatial data is done with the use of pixel aligned priors, used to help stabilize the model in regards to dynamic poses and varying environmental conditions [2]. Incorporation of these parts enables these systems to perform better in terms of generalization and reduces distortion and inconsistency also present in previous generative approaches.

To train the proposed model, data preprocessing is extensive in order to extract accurate features as well as align the geometry. We provide a high resolution 2D image collections annotated with the corresponding 3D pose and geometry. Training Pipeline: This pipeline

involves training based on cropping, resizing and normalizing data for keeping up the scales of the features. To increase the model robustness in different environmental conditions, data augmentation techniques like rotation, flipping, color adjustments are applied. Furthermore, geometric transformation techniques are introduced to increase the pose variations and thus to enhance the adaptability during real time implementation.

The model training process is achieved by the hybrid loss function which discontinues the adversarial, perceptual, and geometry loss to maintain the balance between the realism, texture sharpness, and pose accuracy. We derive the adversarial loss from a discriminator network which attempts to discriminate between real and generated 3D models, so that the generated 3D models have photorealistic textures and facial detail accuracy [1]. Visual discrepancies between generated and ground truth data are minimized with perceptual loss, while geometry loss ensures spatial consistency of pose estimation [2]. The hybrid loss structure in this work improves convergence stability, and thus the performance of generating the dynamic 3D models.

To evaluate model performance under different conditions such as dynamic lighting, multiple poses, and different facial expressions, the experimental setup includes testing the model under these conditions. Dynamic relighting simulations inspired from LumiGAN framework [3] are used to evaluate texture quality for different illuminations. The model is evaluated on both the unseen pose and the generation of consistent structural geometry using the real world datasets [4] which measures sequences of motion and expressive facial data. Furthermore, the real time performance of the framework is measured by recording inference latency while synthesized models should keep visual fidelity with as little delay as possible.

Multiple metrics are utilized to make sure the constraint has meaningful performance evaluation. Image quality and texture realism are quantified using structural similarity index (SSIM) and PSNR. The accuracy in 3D geometry alignment is measured using Chamfer distance, and motion smoothness metrics measure how much our model has stability and manages to avoid motion artifacts in dynamic scenarios. In addition, these evaluation criteria enable to have a complete view regarding the model accuracy, efficiency as well as its adaptability in a real world situation.

We collect data for the experimental evaluation using publicly available datasets from well known benchmarks including SMPL [6] and 3D Morphable Models (3DMM) [7]. It has both pose variants, facial expressions, and high resolution texture maps for model training as well as test. Input video streams from RGB cameras are used for real time performance evaluation, in which 3D models are dynamically synthesized under different environments.

This is further tested by integrating the generated models into augmented reality (AR) environments to assess visual coherence of the models during interactive sessions.

A step forward on the proposed methodology has been to integrate the current advances in generative AI and better architectural designs to enable real time generation of 3D human models. Leveraging volumetric CNNs for exhibiting better structural precision [9], pixel aligned priors for enforcing texture consistency [2] and dynamic relighting capabilities [3], the framework handles the limitations of earlier approaches. Validation from experimental results demonstrate the system's capability in obtaining photorealistic textures, aligned geometry consistently, and generalizing across dynamic conditions on the generated model.

IV. RESULTS AND DISCUSSION

The evaluation of the proposed framework to generate hyperrealistic 3D human models was performed across a number of criteria: visual fidelity, geometric accuracy, its ability to operate in real time and provide an intuitive workflow for the artists and designers. Integration of volumetric CNNs [9] with pixel aligned reconstruction priors [2] in experimental results led to much more realistic and stable generation of models. Finally, while the hybrid architecture was able to solve some of the main problems presented in the existing approaches, such as model consistency across dynamic poses and facial expressions, and varying environmental conditions.

As far as visual fidelity is concerned, the proposed model outperformed existing frameworks namely AG3D [1] and Get3DHuman [2]. AG3D accomplished the synthesis of realistic texture in static poses well, but the performance degenerated significantly under dynamic motion and extreme pose variations. The pixel aligned priors in the proposed method helped to enhance texture sharpness by enabling maps to be stable even during rapid pose transitions. The visual fidelity of proposed method was evaluated experimentally for Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) and 12% and 9% measurement improvements over AG3D [1], Get3DHuman [2] were measured, respectively.

Moreover, dynamic relighting capabilities were provided to the model as an inspiration from the LumiGAN [3] framework to further enhance realism under different lighting conditions. Although LumiGAN achieved better relighting results for static facial models, the proposed method achieved better stability by refining the texture alignment through pixel aligned priors [2]. With this improvement, facial features were kept in their natural structure,

when there were transitions in the dynamic scene. The evaluations in a controlled lighting environment showed that the proposed method outperformed LumiGAN [3] by 15% in terms of SSIM scores on motion intensive scenarios.

Neural avatars were seamlessly synthesized through motion for animated 3D models with the SMPLpix framework [4]. However, traditional methods, for example, SMPL [6] suffered from model distortion during extreme poses leading to visual artifacts and along incoherences. On the other hand, the proposed method utilized volumetric CNNs [9] to help to perceive depth and align features, and with the help of it reduced 18% pose distortion compared to SMPLpix [4], and smoother motion trajectories. In addition, results of Chamfer distance metrics showed better geometric precision in the problematic cases of high pose variation.

One of the major objectives of the proposed framework was to achieve real time performance with minimal effect on visual fidelity. The proposed model performed on average 27 frames per second (FPS) on NVIDIA GeForce RTX 3060 GPU, which shows real-time performance. The performance of this method outperforms AG3D [1] and Get3DHuman [2] which reported 18 FPS and 22 FPS, respectively, under the same conditions. This is due to the optimizations of the volumetric CNN architecture which allows reduced computational overhead without losing model precision. The better performance makes the proposed method fit for interactive applications, such as gaming, augmented reality (AR) and virtual reality (VR).

The proposed method was very adaptable in real world scenarios. The system used RGB video feeds as input to dynamically 3D reconstructed human models with little latency. For all of these live motion capture, virtual production, and interactive avatars, this capability is important. Dynamic relighting features have been integrated to obtain consistent visual quality in changing lighting condition so that the model can be used in more immersive environments.

The proposed method generalizes on a wide range of different input conditions, which contrasts to the limitations that earlier frameworks manifested. Traditional techniques such as SMPL [6] and 3D Morphable Models (3DMM) [7] had an outstanding accuracy in controlled conditions; but their performance was quite unstable in an uncontrolled environment. The proposed model improves robustness on real world datasets with wide facial expressions, pose sequences, background conditions, but by combining volumetric CNNs [9] with pixel aligned priors [2].

However, there are still some limitations to this. The proposed model succeeds in dealing with moderate motion complexity, but it fails in highly dynamic scenes including (but not limited to) sample shown in middle column, in which it fails mostly due to rapid head rotations and/or occlusions. Moreover, it leverages high end GPU resources (may make it

unavailable for low end GPUs and edge computing platforms). Lightweight network architectures and model compression methods will experiment with them for future research if they can solve the problem of efficiency and performance simultaneously.

This study's results demonstrate that generative AI has the ability to synthesize 3D models for real time. The proposed method integrates volumetric CNNs for enhanced structural precision, pixel aligned priors for improved texture mapping as well as dynamic relighting for visual coherence and scales to beyond that of current state of the art. Such advancements have great potential to reduce the challenges of interactivity, and enrich gaming environments as well as digital media creation platforms. This framework also demonstrates real time performance that reinforces the practical significance of utilizing AI driven 3D model generation techniques in immersive AR/VR applications and the applicability of this to increased adoption of these techniques at large.

V. CONCLUSION AND FUTURE ENHANCEMENT

With the proposed framework for creating hyper realistic 3D human models it uses a hybrid architecture that combines volumetric convolutional neural networks (CNNs) [9] with pixel aligned reconstruction priors [2] that show a great improvement over the conventional methods. We present techniques of integration of these techniques, that succeeded in improving visual fidelity, geometric consistency and real time performance. Essentially, the study provided an effective solution to a range of key challenges in dynamic pose handling, texture stability, and computational efficiency as a robust real-time 3D human model generator.

Experiment results demonstrate that proposed method significantly outperform previous frameworks including AG3D [1], Get3DHuman [2], LumiGAN [3]. In addition, the superiority of the framework's realism is shown through visual quality improvement, explained by increased SSIM and PSNR metrics. Further, the dynamic relighting features as in LumiGAN [3] were incorporated to achieve improved visual coherence of the rendered lighting under varying lighting conditions. Unlike previous methods that were unstable at fast pose transitions, the proposed model had its features stay aligned through pixelated priors [2] to create smooth appearance under all environments.

The real time performance of this research was one of the most significant achievements. The inference speed of the model was 27 frames per second (FPS), which is much faster than AG3D [1] (18 FPS) and Get3DHuman [2] (22 FPS). The feature of this

performance improvement can be credited to the optimized volumetric CNN architecture, which leads to an efficient balance between the computational complexity and model precision. In order to run applications such as gaming, augmented reality (AR) and virtual reality (VR) that perform dynamic motion synthesis and user interaction, ... a near real-time performance is essential.

These results have practical implications in several industries where digital content is created (content creation for movies, medical visualization). This framework in AR/VR environments supports interactive and realistic 3D human avatars, which are drawn to the user without breaking consistency and coherency of avatar movement and visual style with the user movements and lighting conditions. Additionally, the RGB video feed 3D models can be reconstructed making the system available for live motion capture, interactive avatars and digital character animation. The model also allows the improved texture sharpness and geometric precision to be applied in realistic simulations where detailed feature mapping is important for improving visual authenticity.

Although promising performance is shown by the proposed method, there are some limitations which can be improved in the future. The model has a notable limitation of reduced stability under extreme motion conditions such as rapid head rotations or fast changing environment, or occlusions. However, as with pixel aligned priors [2], there was improvement during texture alignment during dynamic motion but poor performance when the model had to rapidly change or unpredictably change its 'pose'. The implications of this limitation are that methods for predicting motion or temporal modeling strategies are required that improve motion stability in highly dynamic scenes.

Another one is the computational costs of the model computation. Even though it can run in real time on these high end GPUs like the NVIDIA GeForce RTX 3060, the work is limited by the fact that it relies on these high end compute systems. The remainder of this paper discusses lightweight network architectures and model compression strategies for imposing computational overhead while preserving visual fidelity, leaving the future work for such research. Fortunately, techniques like knowledge distillation, pruning, and quantization could be a solution to improve the efficiency of the model on less powerful hardware platforms.

Additionally, the proposed method synthesizes 3D human models for indoor and controlled environment, but its performance in outdoor or unstructured environment needs to be evaluated. Unreliability of models could possibly be affected by environmental factors such as inconsistent lighting, background noise, and dynamic object interference. More effort will

be made to enable the system to generalize better to make it environmentally robust by incorporating adaptive feature extraction techniques and integrating real time depth estimation.

Another scope for future improvement is in the expansion of the dataset used in the training process. Based on the datasets inspired by AG3D [1], Get3DHuman [2], and SMPL [6], which give high quality pose data but lack diversity in the facial expression, the type of the clothing, the environment conditions, etc., we train the current framework. The rich datasets with diverse ethnicities, multi poses, and none standard lighting scenarios bring the model to be able to generalize better for real world applications.

Furthermore, the practicality of the system will be improved by enhancing the user interaction features of the system. Extension of the framework for creative design applications could be achieved through integration of real-time user customized components like interactive pose alteration or custom model editing based on the user inputs. The system can become more adaptive and capable of refining the generated models using user feedback through introducing adaptive learning mechanisms.

Finally, the proposed method successfully integrates state of the art generative AI, volumetric CNNs [9], and pixelaligned priors [2] so that realistic 3D human models are created with more stable, visually desirable and real time performance. Although existing frameworks such as AG3D [1] and Get3DHuman [2] have shown a strong performance in wellcontrolled scenarios, the proposed approach increases adaptability of the method in dynamic and unpredictable environments. Further future enhancements aimed at improving motion stability, reducing the computational overhead, and widening the training datasets, will incrementally solidify the framework's potential in real world applications and thus make it more applicable in AR/VR, gaming, as well as in the creation of interactive digital content domain.

REFERENCES

- [1] Moshel, Michoel L., Amanda K. Robinsonb, Thomas A. Carlson, and Tijl Grootswagersb. "Are you for real? Decoding hyperrealistic AI-generated faces from neural activity." (2020).
- [2] Pataranutaporn, Pat, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. "AI-generated characters for supporting

- personalized learning and well-being." *Nature Machine Intelligence* 3, no. 12 (2021): 1013-1022.
- [3] Ali, Safinah, Daniella DiPaola, Irene Lee, Victor Sindato, Grace Kim, Ryan Blumofe, and Cynthia Breazeal. "Children as creators, thinkers and citizens in an AI-driven future." *Computers and Education: Artificial Intelligence* 2 (2021): 100040.
 - [4] S. Prokudin, M. J. Black, and J. Romero, "SMPLpix: Neural Avatars from 3D Human Models," *arXiv preprint arXiv:2008.06872*, 2020. [Online]. Available: <https://arxiv.org/abs/2008.06872>
 - [5] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 1–17, 2017. [Online]. Available: <https://doi.org/10.1145/3130800.3130813>
 - [6] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 1–16, 2015. [Online]. Available: <https://doi.org/10.1145/2816795.2818013>
 - [7] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194. [Online]. Available: <https://doi.org/10.1145/311535.311556>
 - [8] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301. [Online]. Available: <https://doi.org/10.1109/AVSS.2009.58>

- [9] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1031–1039. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.116>
- [10] H. Dai, N. Pears, W. Smith, and C. Duncan, "A 3D morphable model of craniofacial shape and texture variation," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3085–3093. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.334>

Citation: Narayana Gaddam. (2022). Generative AI for Creating Hyper-Realistic 3d Human Models. International Journal of Information Technology and Management Information Systems (IJITMIS), 13(1), 137-149.

Abstract Link: https://iaeme.com/Home/article_id/IJITMIS_13_01_013

Article Link:

https://iaeme.com/MasterAdmin/Journal_uploads/IJITMIS/VOLUME_13_ISSUE_1/IJITMIS_13_01_013.pdf

Copyright: © 2022 Authors. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Creative Commons license: Creative Commons license: CC BY 4.0

