iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 10 **Issue:** IV **Month of publication:** April 2022

**DOI:** https://doi.org/10.22214/ijraset.2022.41041

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 10 Issue IV Apr 2022- Available at www.ijraset.com*

# Human Motion Imitation using Generative Adversarial Networks

Sandesh Shinde[1], Pritam Thorat[2], Sagar Yadav[3], Poonam Narkhede[4]
[1, 2, 3, 4]*Department of Computer Engineering, Shivajirao S. Jondhale College of Engineering, University of Mumbai*

***Abstract:*** ***Within a unified framework, we handle human image synthesis, including human motion imitation, appearance transfer, and new view synthesis. It indicates that after the model has been trained, it can do all of these jobs. To estimate the human body structure, existing task-specific techniques mostly employ 2D key-points (position). However, they can only represent location data and have no ability to define the person's unique shape or simulate limb rotations. To untangle the position and form, we suggest using a 3D body mesh recovery module in this study. It may define the customized body form as well as model joint placement and rotation. We present a Liquid Warping GAN technique that propagates source information in both image and feature spaces to the synthesized reference in order to retain source information such as texture, style, colour, and face identity. A denoising convolutional auto-encoder extracts the source characteristics in order to accurately characterize the source identity. In addition, our approach allows for more flexible warping from many sources. A one/few-shot adversarial learning is used to increase the generalization capacity of the unseen source pictures. In particular, it begins by putting a model through a rigorous training process. The model is then fine-tuned in a self-supervised manner by using one/few unseen images to create high-resolution (512x512 and 1024x1024) outputs. In addition, we created the imitation dataset to assess human motion imitation and unique view synthesis. Extensive testing has shown that our approaches work better in retaining facial identification, form consistency, and outfit details.***
***Keywords: Human Image Synthesis, Motion Imitation, Novel View Synthesis, Generative Adversarial Network***

## I. INTRODUCTION

Human image synthesis, which includes motion imitation and new view synthesis, seeks to create credible and photorealistic pictures of humans. It offers a wide range of possible applications in character animation, re-enactment, film or game production, and so on. Given a source human image and a human reference image, the goal of motion imitation is to create an image with the texture from the source human and the pose from the reference human, the goal of human novel view synthesis is to create new images of the human body, captured from various viewpoints. Existing approaches for human motion imitation, for example, maybe divided into two categories: image-to-image translation-based pipelines and warping-based pipelines. The image-to-image translation pipeline learns a person-specific mapping function from the skeleton, dense posture, and parsing result of human circumstances to the image from a video with matched sequences of conditions and pictures. As a result, everyone must train their model from start, and a trained model cannot be transferred to other models. Furthermore, it is not extensible to other tasks, such as appearance transfer. Furthermore, previous methods rely mostly on a 2D posture, a dense pose, and a body parsing result. These approaches only consider layout positions and neglect customized form and limb (joint) rotations, which are even more important in human picture synthesis than layout places. For example, if we use the 2D skeleton, the tense posture, and the body parsing condition to replicate the movements of a small person, the tall one's height and size will unavoidably alter.

Transformation flows may be readily determined by comparing the correspondences between two 3D triangulated meshes, which are more precise and produces fewer misalignments than the prior fitted affine matrix from key points.

## II. LITERATURE SURVEY

Haoye Dong, Xiaodan Liang, Ke Gong aimed to resolve challenges induced by geometric variability and spatial displacements via a new Soft-Gated Warping Generative Adversarial Network (Warping-GAN) [4].

Aliaksandr Siarohin, Enver Sangineto, address the problem of generating person images conditioned on a given pose. Specifically, given an image of a person and a target pose, they synthesized a new image of that person in the novel pose [3].

Liqian Ma, Xu Jia, Qianru Sun et al. proposed the novel Pose Guided Person Generation Network (PG2) that allows to synthesize person images in arbitrary poses, based on an image of that person and a novel pose [2].

Lingjie Liu, et al. proposed a method for developing video-realistic animations of real people that can be controlled by the user. In contrast to conventional human character rendering, they did not require the availability of a production-quality photo-realistic three-dimensional (3D) model of the human but instead rely on a video sequence in conjunction with a (medium-quality) controllable 3D template model of the person [6].

## III.METHODOLOGY

The existing task-specific methods mainly use 2D key-points (pose) to estimate the human body structure. However, they merely communicate position information and have no ability to define a person's particular shape or simulate limb rotations. To try to separate the pose and shape, we suggest using a 3D body mesh recovery module in this paper. It can characterise the individual body shape as well as model joint placement and rotation.

To preserve the source information, such as texture, style, colour, and face identity, we propose a Liquid Warping GAN that propagates the source information in both image and feature spaces to the synthesized reference. In addition, our suggested technique allows for more flexible warping from various sources. To further improve the generalization ability of the unseen source images, we apply human novel view synthesis.

### *A. Generative Adversarial Networks*

GANs, or Generative Adversarial Networks, are a generative modelling approach that employs deep learning approaches such as convolutional neural networks. Generative modelling is an unsupervised learning job in machine learning that entails automatically detecting and learning the regularities or patterns in incoming data such that the model can be used to produce or output new instances that may have been chosen from the original dataset. GANs are an ingenious way of training a generative model by framing the problem as a supervised learning problem with two sub-models: the generator model, which we train to generate new examples, and the model, which attempts to classify examples as either real or fake. The two models are trained in an adversarial zero-sum game until the discriminator model is tricked around half of the time, indicating that the generator model is creating credible instances.

### *B.* Human Novel View Synthesis

The goal of novel view synthesis is to create fresh pictures of the same item or human body from various perspectives. The key step in previous approaches is to use convolutional neural networks to fit a correspondence map from observable views to new views. We used CNNs to predict appearance flow and synthesis new pictures of the same item by copying pixels from a source image based on the appearance flow, and we got good results with stiff things like cars.

### *C. Warping Based Methods*

Recent research is mostly focused on conditioned generative adversarial networks (CGAN) Their main technological concept is to take the source picture and the source posture (2D skeleton) as inputs and use GANs to produce a realistic image using a reference position. The only distinctions between these techniques are in their network topologies, warping algorithms, and adversarial losses. We immediately concatenate the source picture and the reference posture and then create 256x256 using a generator with a coarse-to-fine approach. Some older works substitute rich correspondences between the picture and surface of the human body for the sparse 2D key points. Some suggest a multistage adversarial loss in which the foreground (or various body sections) and backdrop are generated independently. These works are concerned with the method of bending the source characteristics into the target conditions, such as skeleton or parsing. We suggest that a transformation flow be learned from 2D key points and that deep features be warped based on the learned transformations.

### *D. Image Animation*

Traditional techniques to picture animation and video re-targeting were developed for specialized domains such as faces, human silhouettes, or motions, and needed a strong knowledge of the animated object. However, such models are not available in many situations. Image animation may also be viewed as a challenge of translation from one visual domain to another. In order to enhance video translation between two domains, several researchers expanded conditional GANs by integrating Spatio-temporal information. To animate a single person, such techniques need hours of footage of that person labelled with semantic information, and therefore must be retrained for each individual.

## IV. IMPLEMENTATION

We present the system with two inputs viz. a source image and a reference video. The image should be of a person standing still with good lighting and preferably plain background. For better results, we can use two images, a front facing and a back facing image. The reference video can be of a person performing any activities where the camera is still. The inputs are then sent into processing where a skeleton is extracted from the reference video. Once that process is done, the source image is wrapped around the skeleton of the reference video. The unseen pixels are brought together by novel view synthesis. After the entire process is done, we should have a output in the format of a video of the subject from the source image imitating the activities performed by the subject from the reference video.

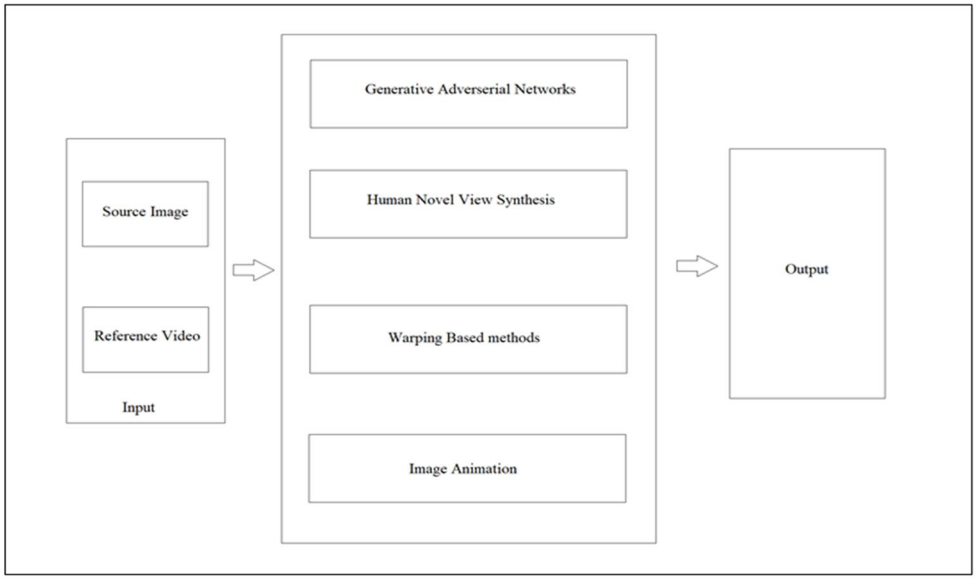



Fig. 1. System Design

As shown in Fig. 1, the body mesh recovery module calculates each image's 3D mesh and produce its corresponding map. The transformation flow is calculated initially by the flow composition module using two correspondence maps and their projected vertices in the picture space. The original picture is then divided into a foreground image and a masked background image. Finally, depending on the transformation flow, it warps the original picture and creates a warped image. The generator in the final GAN module is made up of three streams, each of which creates the background picture, reconstructs the source image, and synthesizes the target image under the reference condition. One advantage of our proposed Liquid Warping method is that it addresses the issue of multiple sources. Also note that Liquid Warping GAN is a unified framework for motion imitation, appearance transfer, and novel view synthesis. Therefore, once we have trained the model on one task, it is capable of being applied to other tasks.

## V. RESULT

Fig. 2. Output

As for data in the dataset, we have videos containing different views of a certain subject performing A-pose, and in the dataset, we render T-pose images with 3D meshes from different viewpoints. Thus, we obtain images of the same person in different views. We randomly sample source images from the testing set of the dataset and change the views. The results are illustrated in Fig 2 and Fig. 3. Our method is capable of predicting reasonable content of invisible parts when switching to other views and keep the source information, in terms of face identity and clothes details, even in the self-occlusion case.

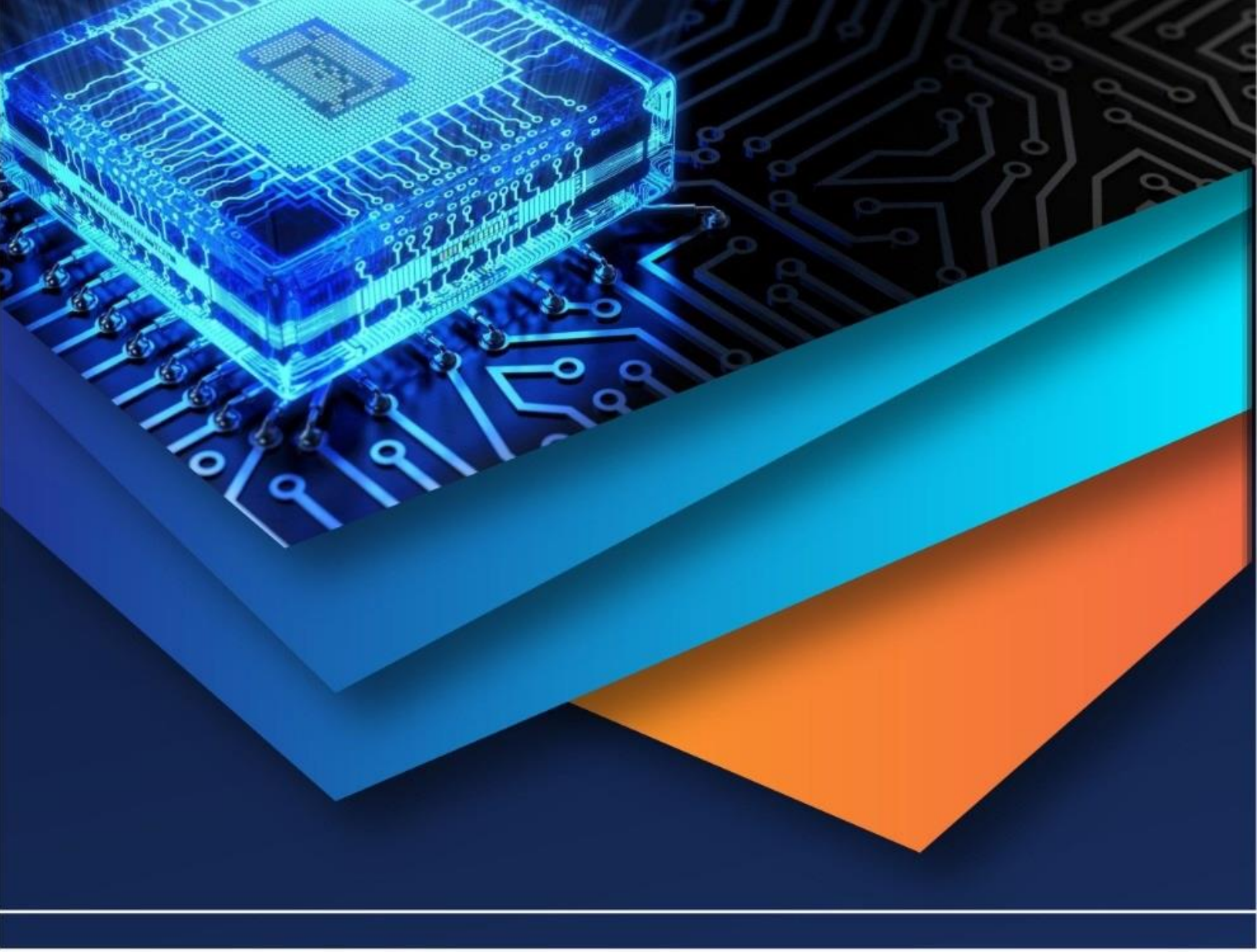
Fig. 3 Output

## VI. CONCLUSION

Our proposed system uses a single framework to manage human motion imitation and new view synthesis. It uses a body recovery module to estimate the 3D body mesh, which is more powerful than the 2D poses. In order to retain the source information, we further developed a new warping strategy, which propagates the source information in both image and feature spaces and supports a more flexible warping from many sources. Besides, with a quick customization, our approach can be extended well when the input pictures are out of the domain of training set and synthesize higher resolution (512x512 and 1024x1024) outputs. Extensive tests demonstrate that our approach out performs others and provide acceptable outcomes.

## REFERENCES

[1] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, “Synthesizing images of humans in unseen poses,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
[2] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, “Pose guided person image generation,” in Advances in Neural Information Processing Systems, 2017, pp. 405–415.
[3] A. Siarohin, E. Sangineto, S. Lathuili`ere, and N. Sebe, “Deformable gans for pose-based human image generation,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
[4] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, “Soft-gated warping-gan for pose-guided person image synthesis,” in Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montr´eal Canada., 2018, pp. 472–482.
[5] N. Neverova, R. A. G¨uler, and I. Kokkinos, “Dense pose transfer,” in European Conference on Computer Vision (ECCV), 2018.
[6] L. Liu, W. Xu, M. Zollhoefer, H. Kim, F. Bernard, M. Habermann, W. Wang, and C. Theobalt, “Neural rendering and reenactment of human actor videos,” ACM Transactions on Graphics 2019 (TOG), 2019.
[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Computer Vision (ICCV), 2017 IEEE International Conference.
[8] T. R. Shaham, T. Dekel, and T. Michaeli, “Singan: Learning a generative model from a single natural image,” in The IEEE International Conference on Computer Vision (ICCV), October 2019.
[9] Y. Li, C. Huang, and C. C. Loy, “Dense intrinsic appearance flow for human pose transfer,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
[10] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, “Video based reconstruction of 3D people models,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY