# Transfer Learning based Parameterized 3D Mesh Deformation with 2D Stylized Cartoon Character

**Sanghyun Byun[1], Bumsoo Kim[2], Wonseop Shin[1], Yonghoon Jung[1] and Sanghyun Seo[2*]**
[1] Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University, Seoul, South Korea
[e-mail: egoist12276@cau.ac.kr, wonseop218@gmail.com, dydgns2017@cau.ac.kr]
[2] College of Art and Technology, Chung-Ang University, Anseong, South Korea
[e-mail: bumsookim00@gmail.com, sanghyun@cau.ac.kr]
[*]Corresponding author: Sanghyun Seo

*Received August 29, 2023; revised October 19, 2023; accepted November 3, 2023; published November 30, 2023*

## Abstract

As interest in the metaverse has grown, there has been a demand for avatars that can represent individual users. Consequently, research has been conducted to reduce the time and cost required for the current 3D human modeling process. However, the recent automatic generation of 3D humans has been focused on creating avatars with a realistic human form. Furthermore, the existing methods have limitations in generating avatars with imbalanced or unrealistic body shapes, and their utilization is limited due to the absence of datasets. Therefore, this paper proposes a new framework for automatically transforming and creating stylized 3D avatars. Our research presents a definitional approach and methodology for creating non-realistic character avatars, in contrast to previous studies that focused on creating realistic humans. We define a new shape representation parameter and use a deep learning–based method to extract character body information and perform automatic template mesh transformation, thereby obtaining non-realistic or unbalanced human meshes. We present the resulting outputs visually, conducting user evaluations to demonstrate the effectiveness of our proposed method. Our approach provides an automatic mesh transformation method tailored to the growing demand for avatars of various body types and extends the existing method to the 3D cartoon stylized avatar domain.

***Keywords:*** 3D Modeling, Mesh Deformation, Computing methodologies, 3D Stylized Avatar, Metaverse Applications

This research was supported by the Chung-Ang University research scholarship grants in 2023 and Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency(KOCCA) grant funded by the Ministry of Culture, Sports and Tourism(MCST) in 2023(Project Name: Development of digital abusing detection and management technology for a safe Metaverse service, Project Number: RS-2023-00227686, Contribution Rate: 100%)

**http://doi.org/10.3837/tiis.2023.11.012** **ISSN : 1976-7277**

## 1. Introduction

Modeling 3D digital humans is extremely useful in multimedia, virtual reality, gaming, and other content areas. Once a 3D human model is constructed, it can be visualized and animated as a digital human for various applications in different fields. With the recent rise of the metaverse, there has been a surge in demand for avatars that can represent individual users in virtual spaces. Avatars are an important means of expressing oneself and acting in a virtual environment, so it is important to generate diverse and distinctive avatars; this highlights the importance of digital human modeling technology in creating avatars. The initial investment required for existing 3D modeling software and hardware is currently substantial. Skilled 3D modelers and animators are essential to creating high-quality 3D human models. The process of creating a 3D human model involves initial model construction and transformation processes such as sculpting and retopology. Each step requires a considerable amount of manual work, resulting in a time-consuming process that can take several weeks or months for a single model. Additionally, if an individual 3D human model needs to be adjusted according to specific requirements, manual modification and reconstruction of the model must be done. This lack of scalability limits the efficiency of the existing 3D human modeling process. Thus, there is a need for improvement to overcome the economic and time-consuming limitations of traditional 3D human modeling.

To overcome these limitations, various approaches to 3D human modeling have been devised. Techniques have been proposed to acquire whole-body data through 3D full-body scanning and use it as the basis for modeling[17][23-24][26]. Such methods provide a high degree of accuracy in creating 3D models of the human form. This 3D scanning method based on scanners has limitations in terms of cost and is vulnerable to noise. Recently, research has been conducted on the automatic generation of 3D digital humans and avatars using deep learning–based technology[19][27-29]. This technology extracts information about the target person from a single RGB or RGB-D image and then creates a 3D human mesh based on this information. These methods not only create realistic 3D humans but also reduce the resource demands of existing processes, thereby increasing their economic efficiency.

However, most existing research has focused on creating human-shaped avatars, whereas there is an increasing demand for avatars with non-human shapes, such as those based on animated characters. Previous studies have mostly focused on reducing reconstruction errors between real people and 3D humans or on increasing realism. Furthermore, in the case of deep learning-based mesh generation methods, they can produce high-quality mesh information, but their utilization is challenging due to the lack of datasets. In conclusion, existing research in this field lacks scalability when it comes to applying it to non-realistic avatars with shapes different from real individuals, such as in metaverses or games. Moreover, most of these studies perform a uniform transformation process based on parameters, which limits the representation of detailed body proportions and can result in inconsistent mesh structures when merged. In contrast to the aforementioned studies, research has been conducted on generating 3D information in unrealistic image domains[3]. However, these studies have limitations as they either require input of unstructured data called sketch lines or can only produce upper body outputs, thus having limited applicability. For this reason, we propose a new method for automatically transforming template meshes using a semantic deformation technique.

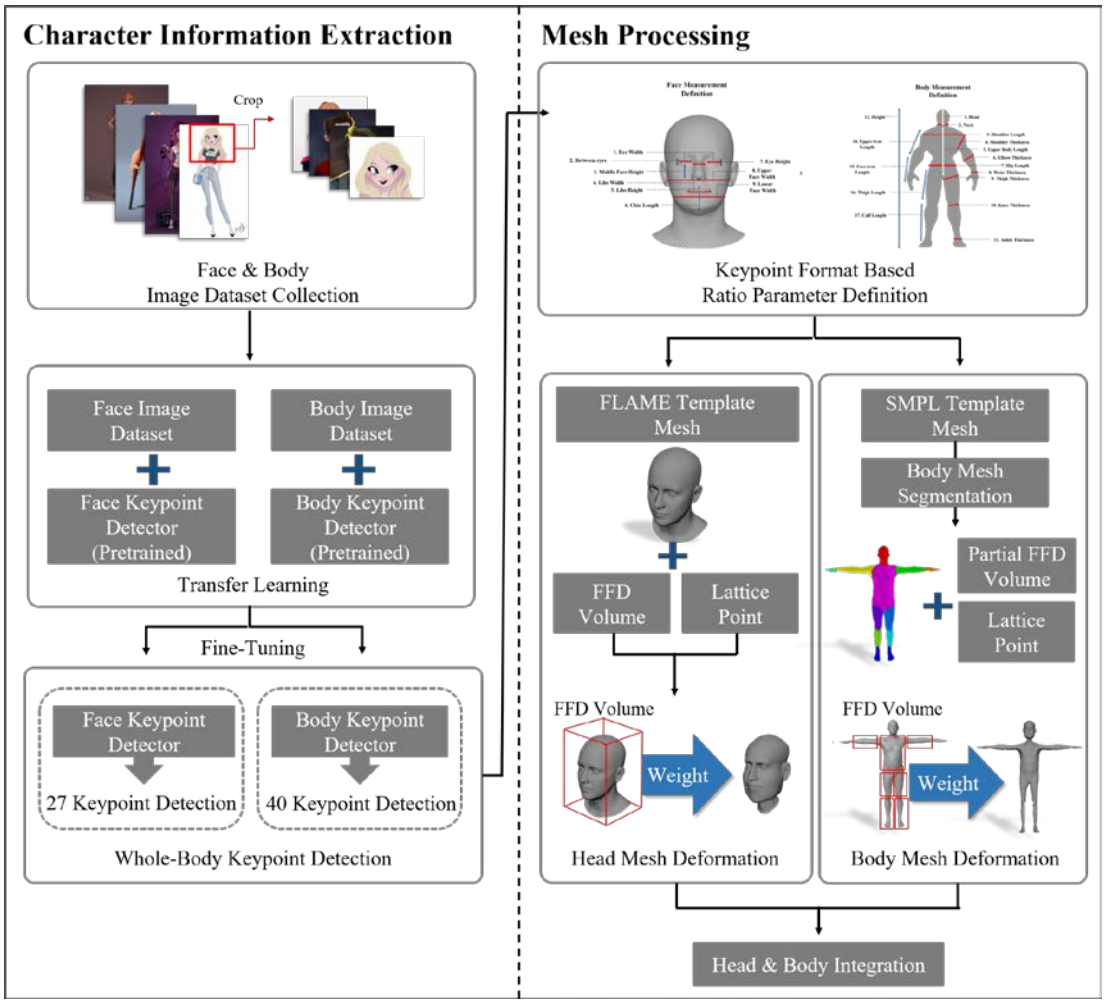


**Fig. 1.** This figure visually illustrates the pipeline of our proposed method. Our approach consists of two major steps: stylized character image information extraction and 3D mesh processing.

In this paper, we propose a novel framework that extracts the shape and proportions of a stylized character from a single image using a deep learning–based method and applies this to perform automatic transformation of human meshes. Key here is that the target for extracting information from the image is a non-realistic illustrated or game character. In essence, we perform non-rigid 3D transformations on an explicit template mesh input that matches the proportions and shape of the reference image character. Building a template-based mesh deformation process provides more freedom from noise compared to methods that estimate or generate meshes. Additionally, by adopting the approach of utilizing well-organized template meshes, as opposed to the 3D meshes derived from existing research, our methodology facilitates easier application and customization of the obtained meshes. At each step, we use a deep learning–based approach to build an objective and automatic process. Additionally, while our methodology focuses on shape transformation, we aim to show in the final stage that it can be easily applied to animation as well. Ultimately, our goal is to present an extended methodology that extracts and utilizes character information from a single RGB image to develop a practical and effective framework for reconstructing avatars with various shapes that are not necessarily human. **Fig. 1** shows the pipeline of the proposed method.

Our work also opens new avenues for further research in creating avatars with diverse shapes and appearances, as well as developing more advanced animation and control methods. Additionally, by utilizing the proposed method, the production process of avatar-based content can reduce the cost associated with the labor-intensive stages of 3D modeling. The key contributions of our proposed methodology can be summarized as follows:

- Our proposed methodology provides an extended approach for representing characters with non-realistic or imbalanced body shapes.
- The proposed method can fundamentally solve the absence of datasets and is free from the difficulties of data construction.
- Our work also opens new avenues for further research in creating avatars with diverse shapes and appearances, as well as developing more advanced animation and control methods. Additionally, by utilizing the proposed method, the production process of avatar-based content can reduce the cost associated with the labor-intensive stages of 3D modeling.

## 2. Related Work

In this section, we report on investigating existing studies on creating 3D human models to support our research. Through this process, we have developed ideas to devise different approaches from existing conventional methods. We also indicate the distinctive nature of our research compared to existing studies. **Fig. 2** summarizes the input and output of the existing researches mentioned in this section, as well as their limitations.

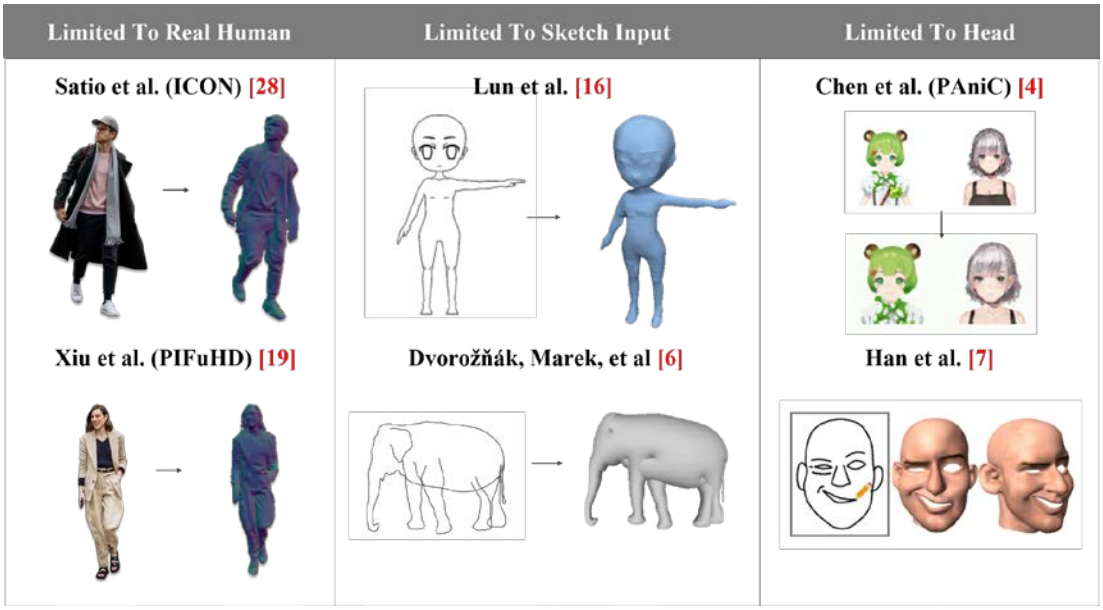


**Fig. 2.** The existing deep-learning methods that have been researched are limited either in terms of producing outputs restricted to real individuals or having limited input and output domains.

3D human modeling and digitalization involve a wide range of topics including human shape and pose estimation, 3D body reconstruction, 3D facial reconstruction, and avatar creation. Moreover, this field aims to accurately obtain high-fidelity 3D models of the human body from a single image or a series of images. Human modeling for widespread application in contexts such as the metaverse and gaming has been researched for decades. Therefore, this section focuses on discussing previous work related to the modeling and deformation of 3D human avatars and examining the progress made in this field over time. We mention the latest developments using advanced techniques for more detailed avatar representation, starting from the early work that laid the foundation for human modeling.

3D human modeling and deformation can be classified into direct model creation, image-based reconstruction methods, and deep learning–based modeling methods. Direct model creation techniques using 3D whole-body scan data were proposed and developed from traditional polygon modeling techniques[1][18][21][26]. The 3D whole-body scan data utilization technique has become capable of generating high-quality data when combined with the development of image capture technology. This has enabled the creation of higher-accuracy 3D models. The extracted 3D scan data can represent human shapes well, and post-processing work using this data is utilized in the creation of digital human meshes. In this context, a method for constructing an actual human model from surface scan data was proposed. Additionally, some studies proposed a framework for creating a consistent mesh structure of multiple 3D whole-body scans, building a database, extracting parameters, and ultimately generating the final model. Although these methods reduced the processing time for constructing and deforming human models using 3D scanners, they did not adequately address the cost issues associated with utilizing scan data.

On the other hand, image-based reconstruction methods involve reconstructing a 3D human shape from information obtained from 2D images. Existing research has utilized image processing technology to extract shape information such as body silhouettes or body measurements as a basis for 3D body modeling[11][25]. This approach requires 2D image data as input, reducing the cost of preparing additional information. However, classical image-based approaches have limitations due to noise and dependence on background environments.

The recent development of deep learning technology has made it possible to extract accurate and detailed information from images. These methods encode the characteristics of the target individual from images and apply them to implicit surface representation techniques to obtain 3D meshes that represent the person's shape well. However, images in the anime or cartoon style domain have more complex and diverse geometric features. Moreover, they are shaded with non-realistic outlines, making it challenging for existing human generators to work seamlessly. These methods can only represent realistic humans and have difficulty achieving partial and detailed variations. In conclusion, while these deep learning-based techniques can be utilized in future digital human creation processes, they lack scalability when it comes to creating stylized 3D avatars that require semantic and uneven transformations. In contrast to that, research has been conducted on reconstruction methods based on sketch-line input from sketch images, which are relevant to generating stylized 3D information[6-7][16]. These researches have built neural networks to learn the underlying 3D representation from a single 2D sketch image or to map it to a 3D model. Additionally, there have been studies on reconstructing stylized 3D heads from anime character illustration images [4]. However, these reconstruction methods tend to rely solely on sketch input or produce limited results, focusing on deriving only the upper body or including only the face. Therefore, a comprehensive methodology is needed to obtain 3D meshes that represent the entire body evenly.

In this paper, we propose a comprehensive methodology for generating 3D stylized avatars. Building upon the research conducted on the upper body, we extend our approach to encompass the entire body, aiming to address limitations observed in previous studies. This extension aims to overcome the constraints of limited styles in input stylized characters and the restriction of input limited to outline images, characteristics evident in previous research. Our methodology employs the most common RGB images as input and can accommodate a wide range of character styles, from 2D illustrations to 3D game characters, ensuring versatility. Furthermore, by adopting a geometric transformation method utilizing template meshes, we maintain mesh quality without degradation. This fulfills a necessary condition for future applications of avatar meshes, namely, the requirement for correct mesh topology to be considered.

## 3. Pre-Analysis

The existing method of creating 3D humans based on images has made significant progress with the development of deep learning technology. These deep learning–based human creation methods can process high-resolution images based on architectures that improve prediction accuracy and obtain detailed meshes. In the prominent studies, however, datasets consisting of rendered 3D meshes of real human shapes are used. Exemplary research in the field of 3D human generation that has provided significant inspiration for subsequent research is the PIFuHD method[19], which introduced implicit functions. In this research, deep learning models were trained using 3D scan datasets incorporating real human body measurements. Acquiring actual scan information for humans is relatively straightforward. However, generating datasets for stylized characters with non-typical shapes presents significant challenges. Unlike humans, where deep learning models can easily generalize from standardized data, the diverse forms of stylized characters mentioned earlier pose limitations in this regard. This analysis has been conducted in various prior studies. The research by Yuda Qiu et al. [30] identified the challenge of mismatch in the reconstruction techniques necessitating accurate registration in the case of non-standardized caricatures and the stylized character domain. This mismatch renders it unsuitable for projection-based applications such as texture restoration and manipulation. This finding underscores the difficulties in dataset construction. Additionally, the study by Wu, Qianyi, et al. [31] mentioned the inherent complexity in the diversity of the stylized character domain. Due to its nature as artistic creations, these characters do not reflect real-world physical environments such as lighting information, making them inherently more challenging for general reconstruction efforts. Therefore, it is not feasible to generate meshes in the stylized character image domain due to the lack of appropriate datasets. We attribute this disparity to the differences in image features between actual humans and the stylized character domain. Images from the real human domain possess continuous features and color distributions, whereas images in the stylized character domain do not exhibit the same characteristics. As a result, as illustrated in **Fig. 3**, we observe the generation of uneven surfaces. We observed the results when applying a representative pre-trained model to the domain of stylized shapes, such as cartoon or illustrated characters. **Fig. 3** shows the results when existing single image–based human generators were applied to the character domain. The models used for analysis aim to create detailed and accurate 3D models of the target human with 2D images as input. The results suggest two major limitations.

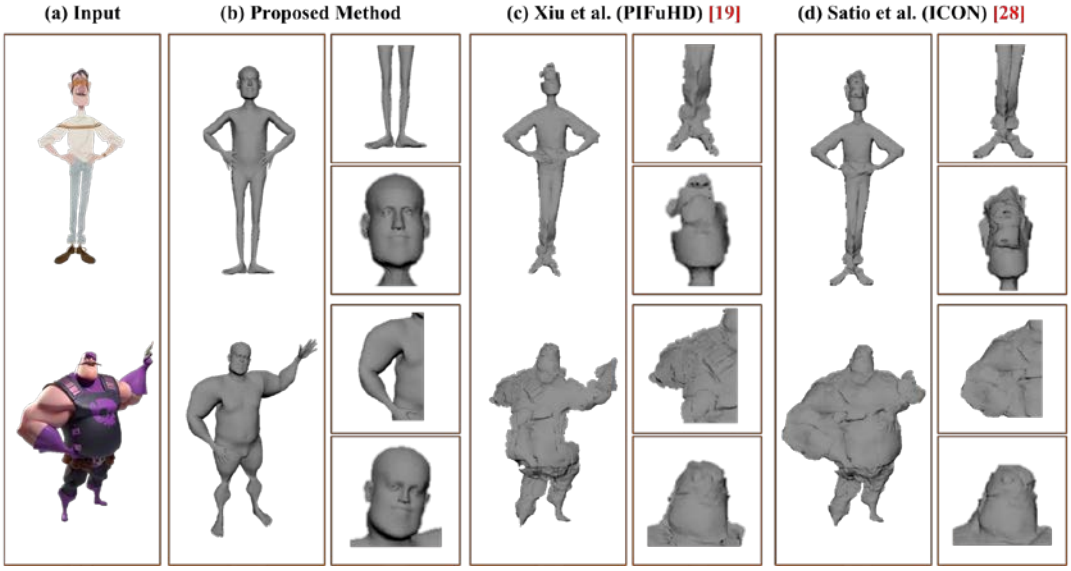


**Fig. 3.** The first and second outputs shown in the figure are the results of the deep learning-based approach using implicit surface representation. As depicted in the figure, these results exhibit mesh merging or vulnerability to noise. On the other hand, the results obtained using the proposed methodology are free from such noise and showcase well-organized meshes.

First, the existing human creation method is insufficient to represent unbalanced or unrealistic shapes. Deep learning–based and implicit surface representation methods show advantages in capturing real human postures and shapes, but they have limitations when applied to stylized character domains with complex and diverse geometric features. Second, we identify topology inconsistency and reduced quality in the results created using existing methods: the generated 3D mesh data shows unstable or broken surface flows. Also, these methods failed to capture facial features and, in some cases, could not perform reconstruction at all. This analytic process allowed us to identify the limitations of deep learning–based 3D human creation methods. We have developed an applied pipeline to assist in the modeling of stylized 3D avatars. In **Fig. 3**, the rightmost image showcases the results of the proposed method. As depicted in the **Fig. 3**, compared to existing approaches, our method provides more stability and obtains meshes that accurately represent the target character. Additionally, it facilitates ease of future transformations and customization for various applications.

## 4. Parameterized 3D Mesh Deformation

We aim to obtain a 3D mesh that can represent characters with non-realistic shapes and proportions based on the input image and template mesh. In our methodology, we identify a series of steps including defining parameters for extracting information from images, transfer learning and adjustment of the keypoint detector, and an automatic mesh processing workflow.

### 4.1 Character Features Extraction from Single Image

We start by building a dataset to train a newly trained keypoint detector, as there is a limited amount of stylized 2D image datasets available, such as anime or cartoon characters. First, we select character samples from various media types, including animation, cartoons, etc. This ensures diversity in body types and proportions, including both male and female

characters. The data includes characters with different styles, including 2D illustrated characters and 3D modeling-based characters.

Next, we acquire high-quality images of each character that are facing forward in a neutral pose, with a clear view of the entire body. For facial data, we crop the face portion from the collected images and resize them for preprocessing the dataset. Specifically, we collected about 3000-character image datasets with various shapes and styles, dividing them into 80% for training and 20% for validation. The images were resized to a fixed resolution of 256x256 pixels. Additionally, we simultaneously constructed a Face Dataset by cropping only the facial parts of the corresponding data.

Annotations for the collected dataset should effectively represent the information in the character images. To achieve this, we adopted a keypoint format commonly used in Human Pose Estimation. For the face, we used an extended format of 27 landmarks based on the AFLW dataset [10] while for the body, we used 40 keypoints in the MPII-TRB [5] keypoint format, which allows for extracting information about skeletal structures as well as contours. **Fig. 4** shows an example of data collected.

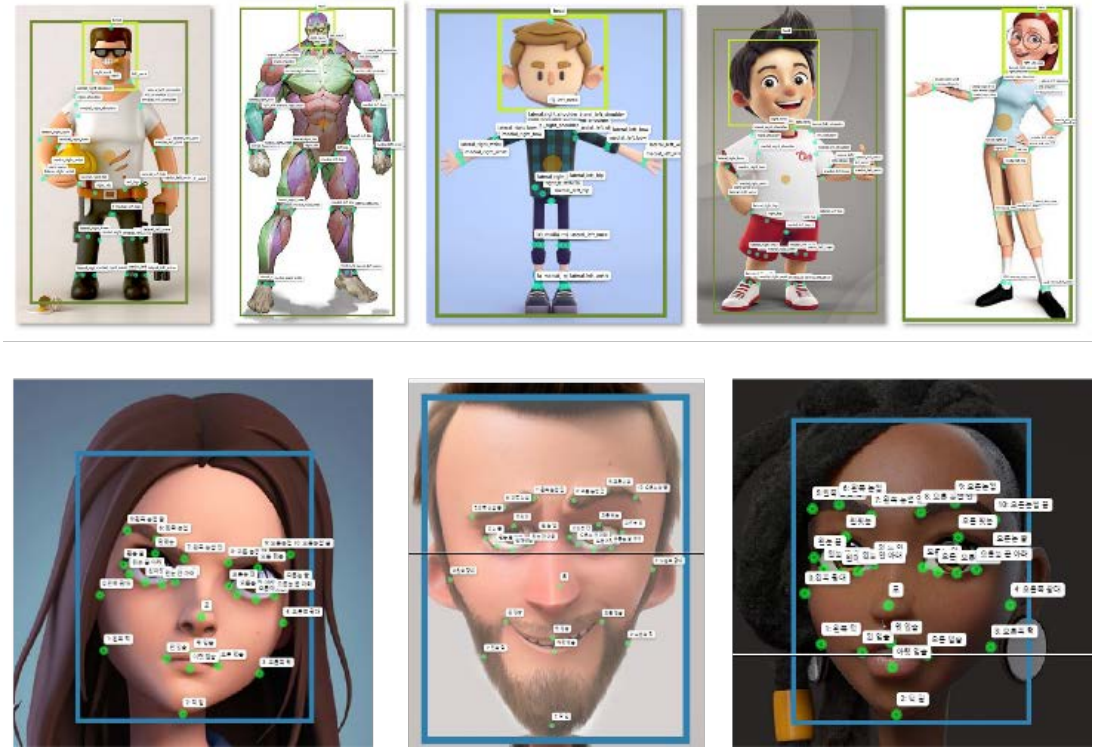

**Fig. 4.** This is an example of the dataset we collected. We gathered a diverse range of stylized character image data. The collected dataset was obtained from an idea-sharing website [34].

Next, we need to perform transfer learning on the collected dataset to train the pre-trained keypoint detectors. **Fig. 5** shows process of training strategy. Based on the strategy shown in the figure, we retrain the original CNN-based deep running model. The weight of the Original Model is used for initialization of the New CNN Model, and only the weight of the keypoint classification layer is updated.

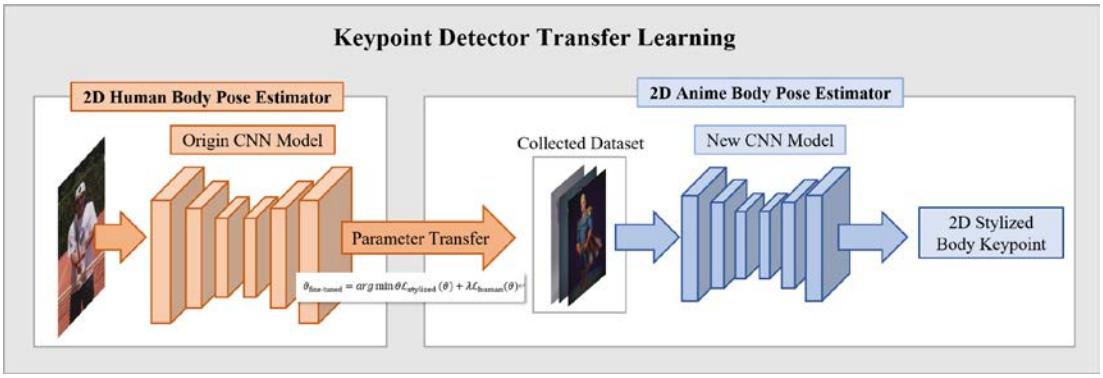


**Fig. 5.** It is a transfer learning strategy using the dataset of the collected Stylized character domain. We use the parameters of CNN-based keypoint detectors learned from Human Domain's dataset to initialize new keypoint detectors.

To perform transfer learning on the acquired dataset, we establish the following equation:

$$\theta_{\mathrm{f}} = arg\min \theta \mathcal{L}_s(\theta) + \lambda \mathcal{L}_h(\theta) \tag{1}$$

In (1), $\theta_{\mathrm{f}}$ represents the parameters of the model after fine-tuning. These parameters are learned to better suit the stylized cartoon character pose estimation task. $L_s(\theta)$ is the loss function associated with the task of estimating poses for stylized cartoon characters using the model's parameter, $\theta$. This loss function quantifies the mismatch between the model's predictions and the ground truth poses of cartoon characters in your training data. The fine-tuning process aims to minimize this loss, effectively improving the model's performance on the cartoon character task. $L_h$ is the loss function associated with the pre-trained pose estimation model using the parameter, $\theta$. This loss measures how well the pre-trained model performs on its original task that human pose estimation. $\lambda$ is a hyperparameter that controls the balance between the two loss terms. It determines the trade-off between fitting the model to the new cartoon character pose estimation task and retaining the knowledge from the pre-trained model.

To find the best-performing detector, we trained detectors with three different architecture backbones: ResNet[8] , SCNet[14], and HRNet[22]. Using the collected dataset, we performed transfer learning and obtained models that output 27 keypoints for the face and 40 keypoints for the body. We updated the weights of the last two layers and fine-tuned the models to align with the Face and Body keypoint outputs. The models were trained with a batch size of 32 for 210 epochs. During this process, we monitored the training loss and validation loss to ensure that the models did not overfit the training data. For the body keypoint detector, we evaluated performance using the percentage of correct keypoints (PCK) metric with thresholds of 0.5 (PCKh@0.5) and 0.1 (PCKh@0.1). The PCK metric measures the percentage of predicted keypoints within a specific distance threshold from the ground truth keypoints. We evaluated the face landmark detector using the normalized mean error (NME) metric. The NME metric measures the average distance between predicted and ground truth keypoints and is normalized by the interocular distance. We present the evaluation results from various architectures in **Table 1**.

**Table 1.** Simple results of body & face keypoint detector

| Body Keypoint Detector | | | Face Keypoint Detector |
|---|---|---|---|
| Arch | PCKh@0.1 | PCKh@0.5 | NME |
| ResNet-50 | 0.3468 | 0.8815 | 0.054 |
| ResNet-152 | 0.3482 | 0.8856 | 0.052 |
| SCNet-50 | 0.3353 | 0.8794 | 0.056 |
| SCNet-101 | 0.3471 | 0.8832 | 0.053 |
| HRNet-w32 | 0.3557 | 0.8885 | 0.054 |
| HRNet-w48 | 0.3665 | 0.8902 | 0.049 |

The simple results for the body detector show that the HRNet-w48 architecture model achieved the highest performance. Similarly, for the face detector, the HRNet-w48 architecture-based model demonstrated good performance. We decided to use HRNet as the backbone for keypoint detection, as it showed the best performance during this process. **Fig. 6** visualizes the estimated keypoints for the face and body in a sample image.

**Fig. 6.** This figure shows the results of keypoint detection on three test images. Keypoint detection is well performed in the stylized character image.

## 4.2 Mesh Deformation Process

In the previous step, we extracted keypoints that provide information about the character's shape. The purpose of this step is to perform automatic deformation of the template mesh based on the extracted information. To achieve this, we defined standard shape parameters for the character. These parameters are based on the 27 facial keypoints and 40 body keypoints for stylized character measurements. **Fig. 7** shows each defined parameter. We perform a

method for automatically deforming the mesh. This process includes 3D template mesh selection, Euclidean distance–based parameter calculation, mesh segmentation, deformation points, and weight derivation. We utilize standard forms of template meshes that are highly compatible with modern graphics pipelines and commonly used in body reconstruction research. Specifically, we use the body model of SMPL (Skinned Multi-Person Linear) [15] and the head model of FLAME (Faces Learned with an Articulated Model and Expressions) [13]. To perform the deformation, we remove the existing parameters of these models and apply the transformation based on the newly defined standard parameters. This allows us to adapt the template meshes to the specific shape and style of the character based on the extracted information from the keypoints.

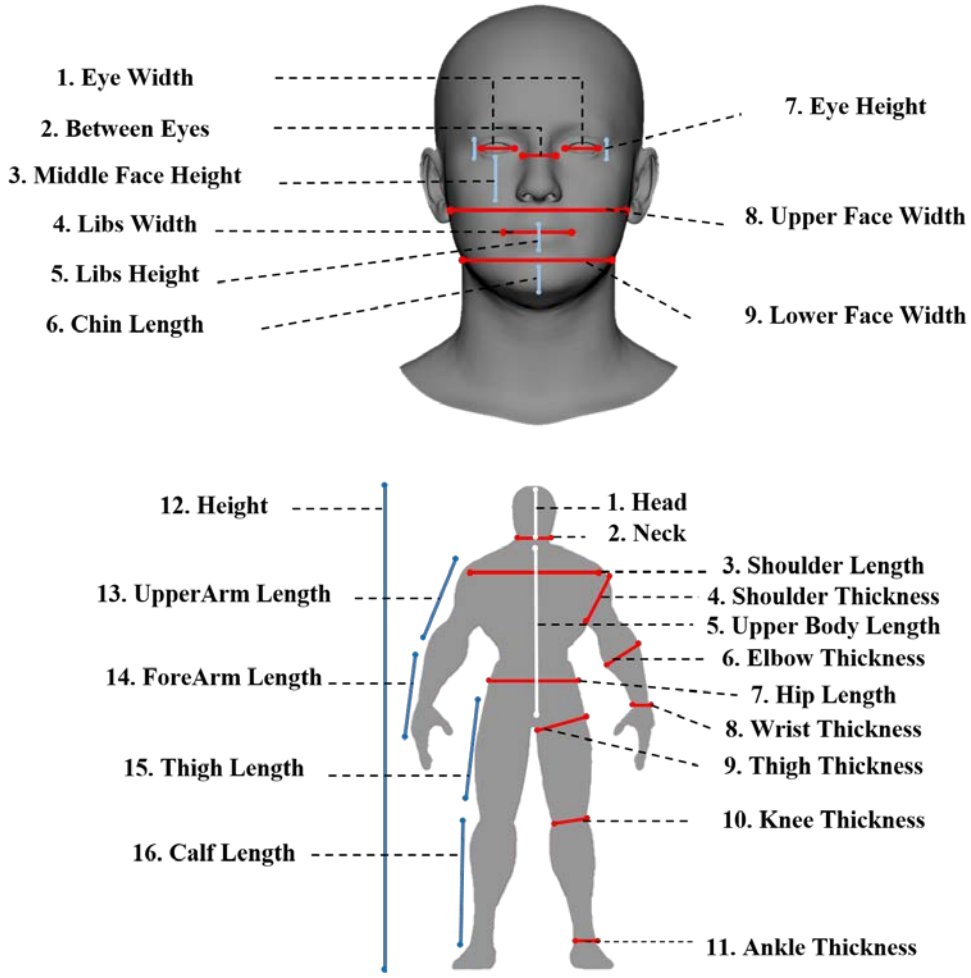


**Fig. 7.** The parameters shown in the figure are defined based on the stylized character body keypoint format.

The first step in calculating body proportions is to compute the distances between key points using Euclidean distance, namely the straight-line distance between two points in Euclidean space. We use the coordinates of the annotated keypoints in the dataset to calculate the Euclidean distance between them. Key points can include joints, facial landmarks, and

other important body parts. To obtain a set of distances, we calculate the distance between all pairs of keypoints. Next, we calculate the relative proportions of various body part lengths based on head length. The head is often used as a reference to determine body proportions in character design and is calculated by measuring the distance between two key points such as the top of the head and the chin. Using this length as a reference, we calculate the relative proportions of other body parts. For example, the length of the torso can be 1.5 times the head length, the length of the legs can be 3 times the head length, and the length of the arms can be 2 times the head length. These proportions may vary depending on the character's style and genre. Additionally, the methodology applied to the body is also applied on an equal basis in facial proportion calculations. Based on the head length, we obtain relative proportions such as eye width and height and nose length. Calculations of the Euclidean distances between keypoints and the relative proportions of other body part lengths based on head length can be used to derive important parameters in character design, such as body part lengths, body proportions, and overall character shape. These parameters can ensure consistency in character design and can be used to create visually appealing and believable characters.

We next semantic deform the input 3D template mesh. To do this, we need to automatically segment this mesh. Our methodology should be applicable even when not using the explicit template mesh employed in this paper. Therefore, it is necessary to automatically perform the task of segmenting body parts by taking the template mesh as input. We use graph convolutional networks (GCNs) to automatically segment the 3D human mesh into different parts. GCN is a type of neural network that operates on graph-structured data. In this methodology, the 3D human mesh is represented as a graph; each vertex represents a 3D point of the mesh, while the edges represent connections between these points. To perform body part segmentation for semantic mesh deformation, we utilized a GCN-based deep learning model. We used the MPI-FAUST dataset[2] for training the model, dividing it into 80% for training and 20% for validation. We then applied the trained model to the template mesh to obtain the 3D coordinates of the vertex sets for each body part. We use the trained GCN to classify each vertex into one of the body part classes.

The method of controlling mesh deformation entails a proposed methodology to perform semantic deformation of the 3D human mesh using 3D free-form deformation technology. The goal is to deform the body and facial meshes while preserving the overall structure and shape of the original mesh. Free Form Deformation (FFD) algorithm[20] is used in the mesh deformation process. FFD is a widely used method in computer graphics for modeling non-rigid deformations, allowing flexible and efficient control of the deformation process. By applying semantic FFD techniques to the template human mesh, we can achieve semantic shape deformation while maintaining a human-like underlying structure and kinematics. This not only enables the creation of non-human avatars but also allows for more intuitive and flexible control of body shape and pose compared to existing methods. First, for the body, we calculate the 3D deformation volume that best encloses the vertices of a segmented semantic mesh subset, and then determine the deformation weights based on the distance between the subset and the remaining mesh. **Fig. 8** illustrates qualitative examples of this semantic deformation process. Semantic deformations in the arms and legs are achieved by specifying FFD volumes.

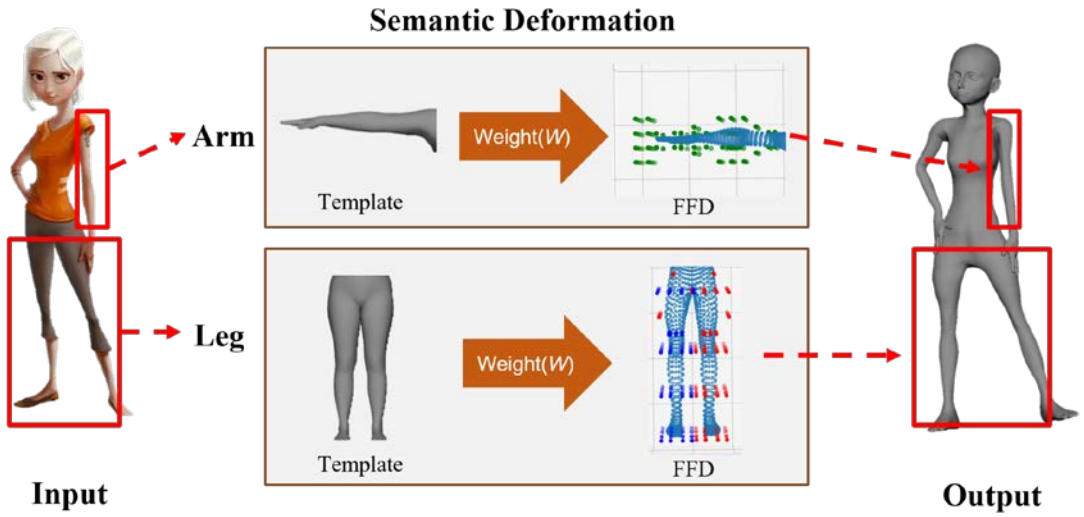


**Fig. 8.** This figure illustrates the process of Semantic Deformation through Free-Form Deformation (FFD). It is the example of deformations in the arms and legs.

We calculate the 3D free-form deformation volume that best encloses the semantic vertex set and use this to deform the subset while preserving the overall structure and shape of the original mesh. To calculate the deformation weights, distance information based on the ratio between the target character and the template mesh is required. For this purpose, the distance from the skeletal key points of the character to the contour key points of each body part is considered, and the distance from the body joints of the SMPL template mesh to the segmented body part mesh is measured at the same angle:

$$D(i) = |D_t(i) - D_s(i)| \tag{2}$$

In (2), $D_t(i)$ represents the distance from the skeletal key points to the contour key points of body part $i$ in the target character and $D_s(i)$ represents the distance from the body joints of the SMPL template mesh to the segmented body part mesh for body part $i$, measured at the same angle. Subsequently, we calculate the deformation amount using the following equation based on the derived weights:

$$V(x,y,z) = \sum_i w_i(x,y,z) P_i(x,y,z) \tag{3}$$

In (3), $V(x,y,z)$ is the deformation volume, $w_i$ is the deformation weight at the point *(x,y,z)*, and $P_i$ is the deformation control point in the form of a 3D bounding box. For the face, we calculate the FFD volume enclosing the entire head, just as for the body. Additionally, in this process we determine the number of control points that can effectively control the shape and proportions of the 3D face. The control point positions of the FFD volume should be well fitted to control the height and width of the eyes, nose, and mouth. We can then deform the facial features of the target character according to the relative proportions of the facial area based on the height of the 3D head mesh, which are used to calculate the deformation weights. Furthermore, our proposed method of semantic mesh deformation control is expected to achieve free and flexible deformation even when the target character's body is asymmetric.

**Fig. 9** and **Fig. 10** briefly show how the mesh is deformed in the proposed method.

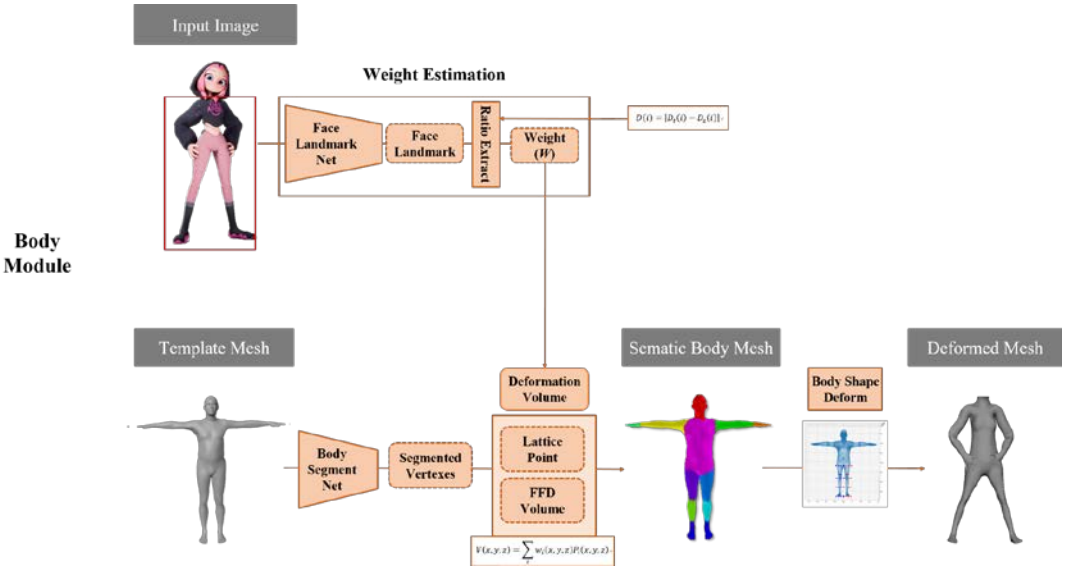


**Fig. 9.** The results derived by applying the deformed volume after mesh segmentation in the Body Template Mesh.

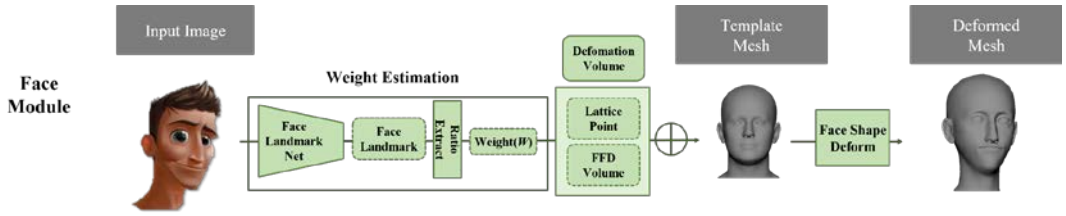


**Fig. 10.** The results of performing the Head Template Mesh Deformation using the information of one sample Face image.

The previously deformed SMPL and FLAME models are separate 3D objects. To obtain an integrated 3D human mesh, we align the positions of the neck vertices of the SMPL mesh and the FLAME mesh on the same line. Then, we unify the positions of vertices within a threshold distance to the median value. We add facial information between the aligned vertices to obtain a single connected mesh. Through this process, we can obtain a whole-body mesh combining FLAME and SMPL models.

## 5. Experimental Results

We have implemented the proposed methodology in a series of step-by-step processes, including character image body dataset collection and keypoint detection as well as human mesh semantic deformation. Through this, we extracted keypoints from various input images of anime, game, and illustration characters, and calculated proportions and shape parameters. Subsequently, we were able to generate 3D meshes of avatars with shapes different from those of real humans. We have determined the following scenarios to demonstrate our research results: 1) Qualitative Assessment of the deformed body mesh: We compare the reconstructed

3D mesh using our methodology with the input image. Through the results, we can judge the expression and quality of the 3D mesh shape. 2) User study of 3D model: We conduct this process to quantitatively evaluate the resulting body mesh. We carry out user evaluations on the final resulting mesh to demonstrate its quality. 3) Extension for avatar application: We showcase the utilization of the created 3D data. By demonstrating 3D rendering within a virtual space, we prove the potential for its application in the avatar content.

### 5.1 Qualitative Assessment of 3D Mesh

We evaluate whether the obtained 3D mesh shape and proportions accurately represent the character in the reference image. To do this, we present qualitative results through a comparison between the mesh and the image. This includes applying the 3D mesh deformation to match the specific pose of the character in the 2D input image. We present qualitative results for our deformed body mesh methodology, focusing on two main aspects: pose matching and reposing. To compare the input image with the 3D mesh, it is necessary to align the target character pose. We are remapping the Pose Parameter in the 3D software. This approach allows us to compare the input image and the generated 3D mesh more accurately. We performed reposing to evaluate the quality of the deformed body mesh when applied to different poses. **Fig. 11** and **Fig. 12** show the results. **Fig. 11** depicts the output when a male character image was given as input, while **Fig. 12** shows the mesh output obtained when a female character image was provided.

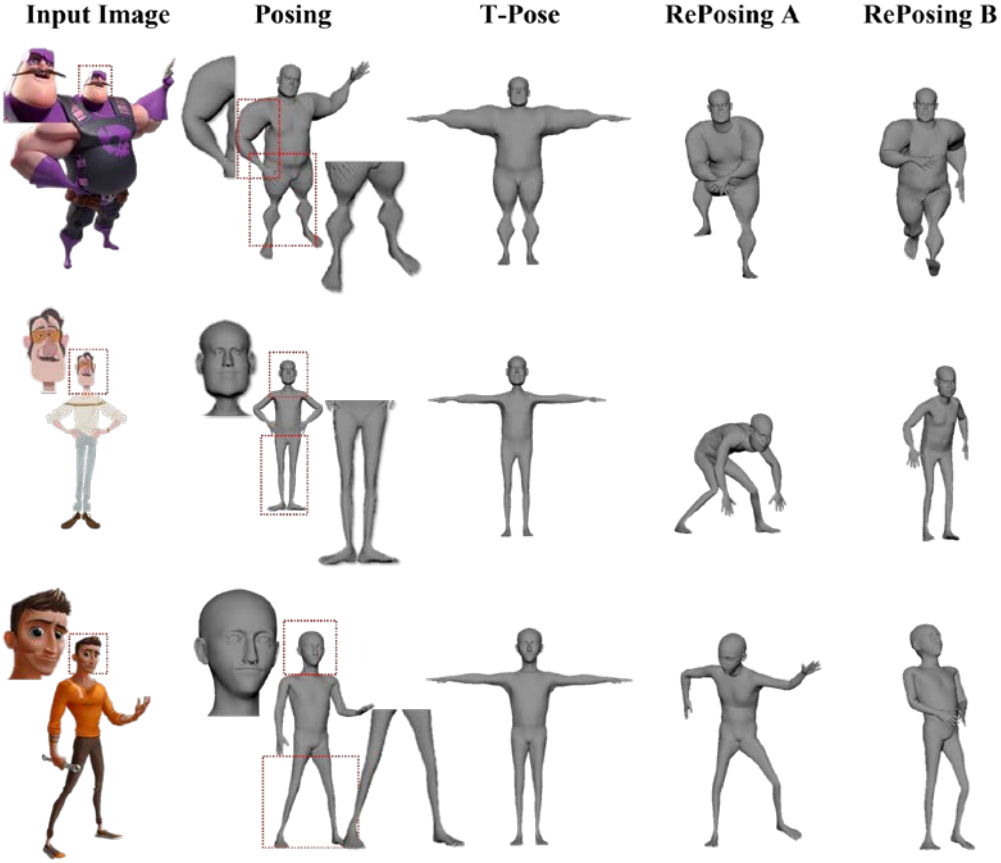


**Fig. 11.** Results of the method using male character image input.

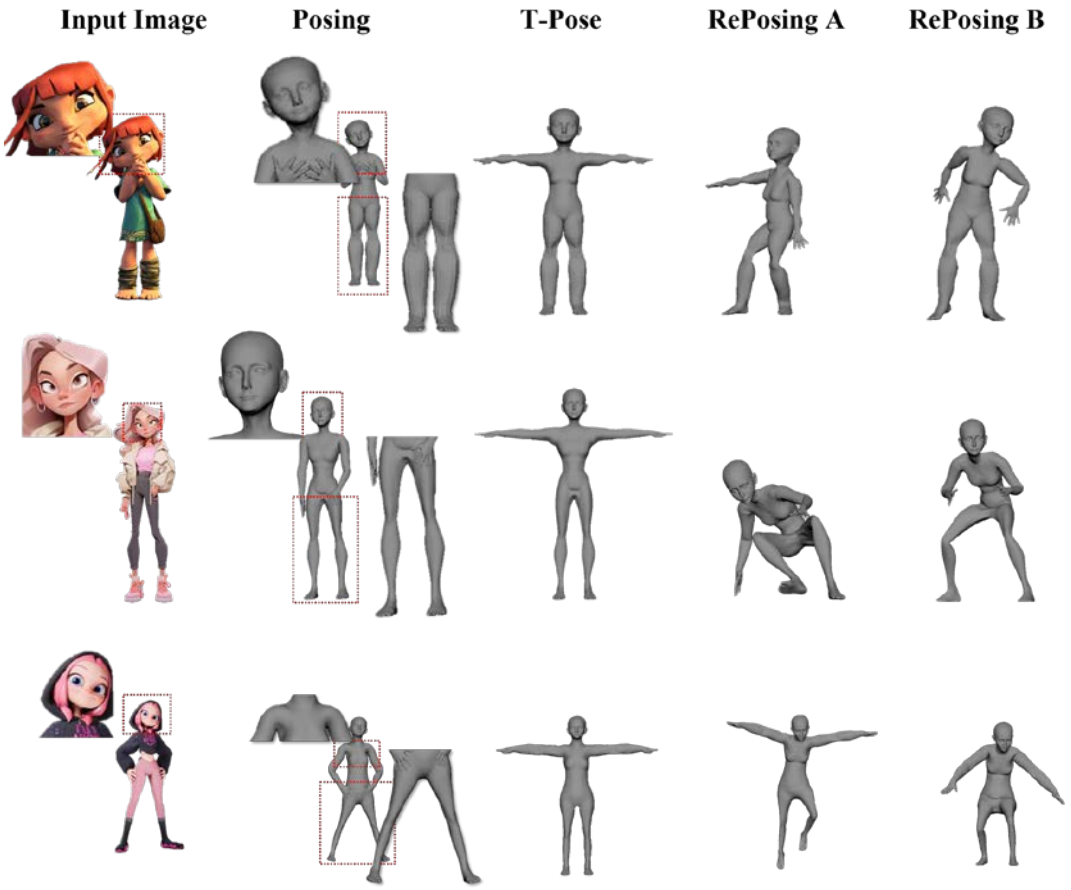


**Fig. 12.** Results of the method using female character image input.

Each body mesh has been deformed to match human proportions, such as face length and limb length. Additionally, this process resulted in smooth deformations without the occurrence of vertex entanglement. As seen in the **Fig. 11** and **Fig. 12**, the derived 3D mesh can accurately represent shapes with imbalanced or unrealistic human proportions. Consequently, we generated a 3D body mesh that maintains the shape and proportions of the character in the input image through our proposed method.

## 5.2 User Study

To validate our methodology quantitatively, we conducted a user study with 36 participants to evaluate the quality and usefulness of the results. Participants were selected from diverse backgrounds such as computer graphics, animation, gaming, and visual arts. Each participant was presented with a set of generated unrealistic 3D human meshes and asked to evaluate the models based on various evaluation items. **Table 2** shows the details of the evaluation questions. The main distinctions in the evaluation questions are visual elements and utility, with three sub-questions for each distinction. We consider a score of 5 or higher as a positive evaluation and 3 or lower as a negative evaluation. In Q1 of **Table 2**, we asked about the consistency between the derived 3D mesh and the reference image character. In Q2 and Q3, we asked about the visual and geometric quality of the mesh respectively. Q4 asked about the

suitability of the proposed method's results for future avatar-based content, and Q5 and Q6 asked about the usefulness in the production process and the feasibility of motion application respectively. We then asked for additional opinions to explore elements that should be addressed in our subsequent improvement work.

**Table 2.** Question Design Table

| Main Category | Sub Category | Question Coding | Question |
|---|---|---|---|
| Visual Element | Consistency of Deformation | Q1 | Is the shape of the 3D mesh consistently represented by the reference image character? (1: Very inconsistent; 7: Very consistent) |
| | Visual Quality | Q2 | Compared to the original template mesh, is the visual quality of the mesh after deformation superior? (1: Very inferior; 7: Very superior) |
| | Geometric Quality | Q3 | Do you think the quality of 3D mesh information such as mesh topology, surface flow, and surface smoothness shown in the results after deformation is excellent? (1: Very inferior; 7: Very superior) |
| Utilizat-ion | Suitability | Q4 | Judging from its appearance at the application stage, is the resulting 3D mesh suitable as a digital human model for use in metaverse, games, etc.? (1: Very unsuitable; 7: Very suitable) |
| | Usefulness | Q5 | Judging from its appearance at the application stage, do you think the resulting 3D mesh can be used effectively in the avatar creation process? (Feasibility in the process) (1: Very unhelpful; 7: Very helpful) |
| | Motion Applicability | Q6 | Judging from its appearance at the application stage, do you think the resulting 3D mesh is appropriately prepared for animation with proper rigging and deformation features? (Feasibility in the process) (1: Very inappropriate; 7: Very appropriate) |
| Opinion | Future Suggestion | S1 | Please leave any additional opinions on aspects you would like to see improved in the future. |

Most respondents chose a score of 5 to 7, which corresponds to a positive evaluation in the similarity category of visual elements after mesh deformation. This indicates that the resulting mesh from our method consistently represents the shape and proportions of the original image. Respondents mostly chose a score range of 5 to 7 in the visual quality and geometric quality items of the mesh, with few negative responses. Moreover, a significant number of participants gave positive evaluations in questions about the feasibility of the results. Specifically, the questions received 75%, 72.2%, 75%, 75%, 69.4%, and 83% positive evaluations, respectively.In addition to the quantitative evaluation, participants provided feedback for future research. Some common themes that emerged from the feedback include the following: 1) Participants greatly appreciated the visual and diverse appeal of the generated non-realistic 3D human meshes. 2) Some participants suggested that certain parts of the mesh should have more stylistic consistency or better control over the degree of shape deformation. 3) Additionally, there were opinions that improvements to the mesh details were needed. **Fig. 13** and **Fig. 14** show graphs of the user survey results and average scores.

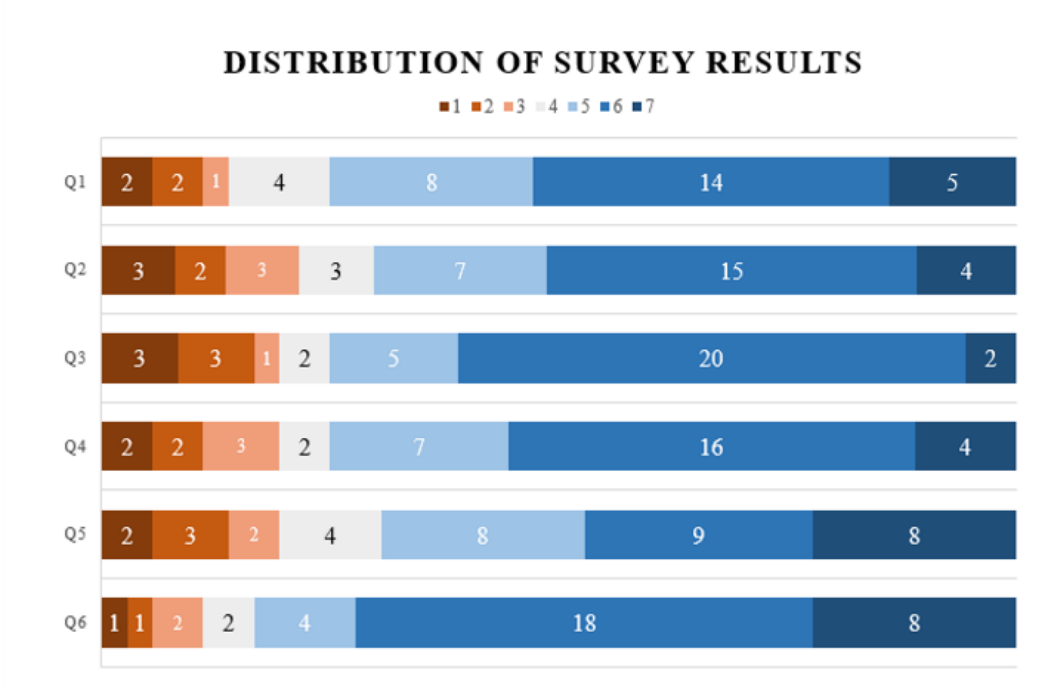


**Fig. 13.** Survey results: Schematic of the frequency of responses as an accumulated graph, (bottom) the average score.

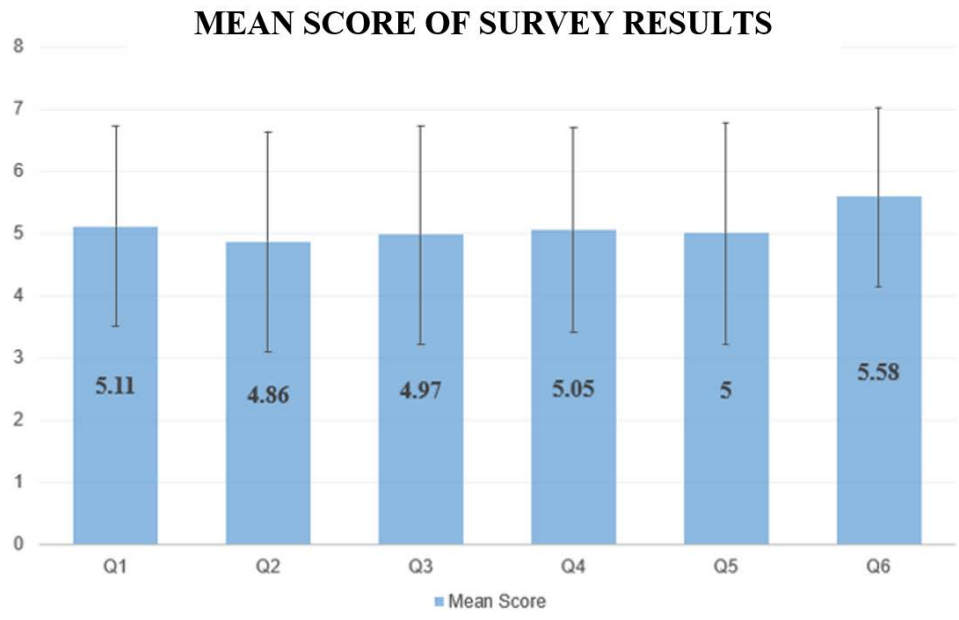


**Fig. 14.** Survey results: Average score.

### 5.3 Extension for Production

In this section, we present the applications of the deformed and reconstructed 3D mesh models using the proposed approach. Our goal is to show that the resulting 3D models can be used to

create more vivid and diverse avatars in video game and virtual reality content. The mesh results can be applied to content in 3D software through re-creation. When aiming to generate avatar meshes at the same level using traditional methods, issues like surface gaps and uneven topology often lead to reduced scalability, requiring extensive time for retouching and modifications if one intends to utilize the results. In contrast, meshes obtained through our proposed method have smooth topology, eliminating issues with mesh integration. Furthermore, there are no limitations on input, and it is possible to obtain full-body meshes, including the head. This convenience in avatar production makes our method highly practical for future applications. These reasons are explicitly illustrated in this section through figures. **Fig. 15** shows a scenario where our methodology is utilized in the creation of avatars.

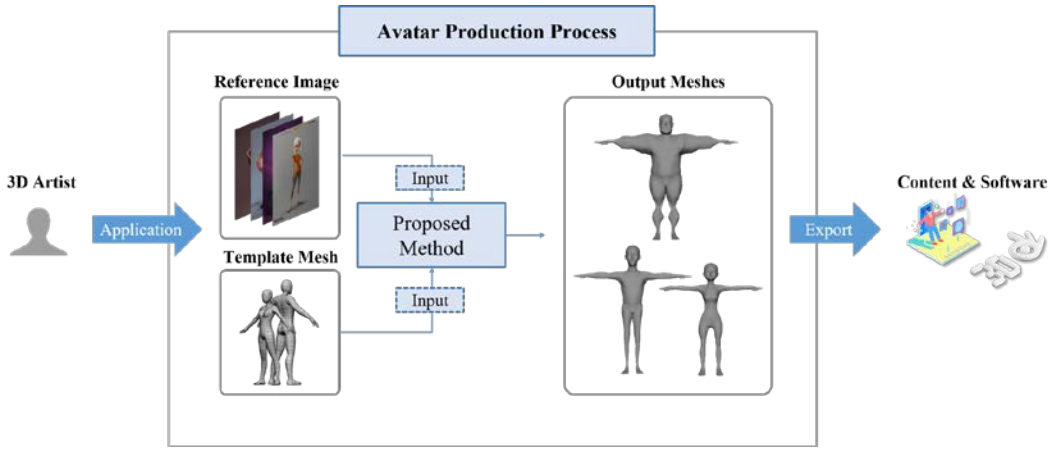


**Fig. 15.** Our method can be utilized by artists aiming to create 3D avatars. By providing reference images and a template mesh for the desired transformation, the program can be extended to generate the transformed resulting mesh. This can be employed in content creation or within 3D software

The final 3D avatar was rendered to be visually presented in a virtual space. Unreal Engine 5 was used for rendering. Furthermore, a mesh deformed by our methodology can be utilized as a base mesh in the human avatar modeling process. To verify this claim, we applied our approach in 3D software. **Fig. 16** and **Fig. 17** show examples of the extended application stage.

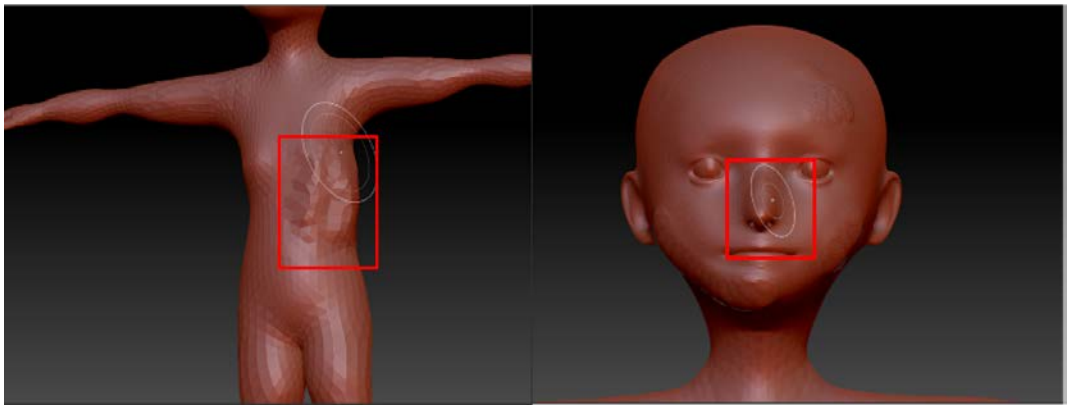

**Fig. 16.** Example of using the resultant mesh with our methodology in 3D avatar content. Figure shows application in Zbrush.

**Fig. 17.** Example of using the resultant mesh with our methodology in 3D avatar content. Figure is rendered on Unreal Engine 5

The 3D human mesh obtained through our methodology can be used as an initial model without quality degradation and offers benefits in avatar content production. Additionally, the meshes obtained through our method maintain their quality even when motion is applied. **Fig. 18** shows the appearance of the mesh when motion is applied.

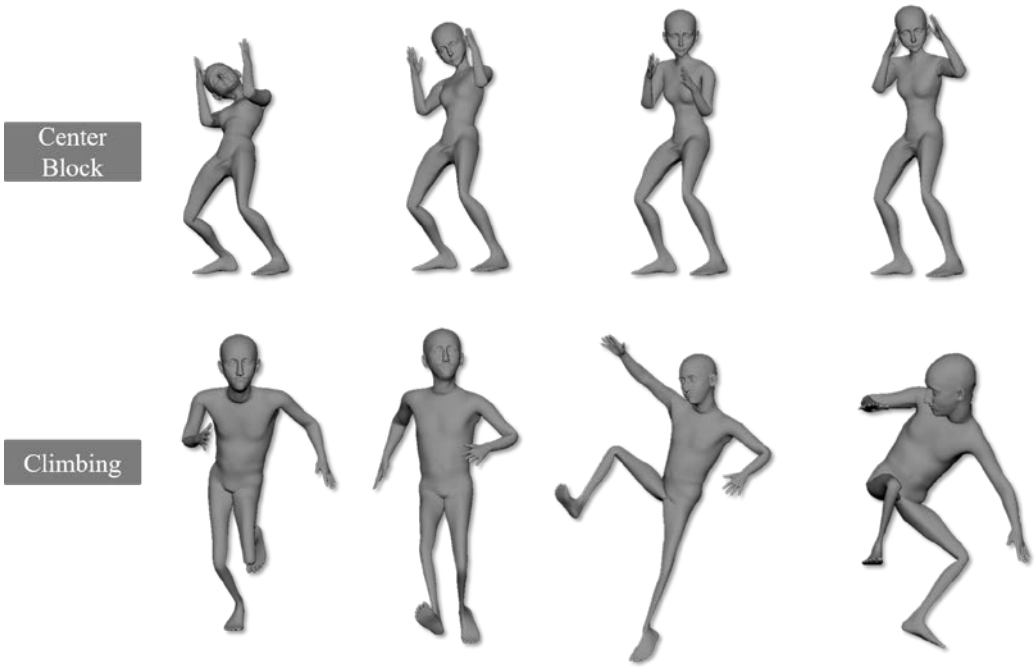


**Fig. 18.** Motion was applied to the two resultant meshes.

In summary, the method proposed in our paper generates full-body 3D avatar meshes representing stylized characters. This has been visually proven to be appealing through user evaluations. Furthermore, unlike meshes generated using traditional methods, these meshes are free from topological errors or inaccuracies. This characteristic is a significant advantage in practical applications. However, our evaluations have revealed certain limitations of the proposed method. It lacks the ability to represent detailed elements such as clothing and accessories, and inaccuracies in the results of 2D keypoint detection can lead to incorrect outcomes.

## 6. Conclusion

In this paper, we proposed a new framework for automatically deforming 3D human avatars of various shapes and proportions using deep learning-based keypoint estimation and 3D freeform deformation. This approach allows us to create avatars that deviate from standard human shapes, a new approach that differs from existing automatic generation methods focused on real people. Our method utilizes transfer learning of a deep learning-based keypoint estimation model to obtain body shape and proportion information of animation characters from a single input image without additional elements. We then reconstruct the avatar reflecting the character's body and facial information using mesh deformation techniques. By comparing the input image and the reconstructed avatar, we demonstrated the efficiency of our approach and its applicability in various virtual content applications.

The reconstructed mesh created by our method can serve as the starting point for 3D character body modeling, simplifying the process and improving scalability for creating diverse avatars. This contributes to enhancing the shape freedom of 3D humans in response to the demand for various stylized character 3D avatars. In addition, we fundamentally solve the problem that existing methods are difficult to use due to the absence of stylized 2D and 3D datasets. Our proposed framework is flexible, has a high degree of freedom, and is easy to control, allowing a wider range of users and developers to access it. This opens new possibilities for creating unique avatars that cater to diverse tastes and preferences, contributing to a more inclusive virtual environment.

Our work makes it possible to obtain avatars of various shapes, but there are limitations in fine expressions such as intricate wrinkles and clothing. Quality degradation occurs because, if the extraction of the keypoints that serve as the basis for information fails, it affects the entire framework. To improve these aspects, future work should explore improving the accuracy of the keypoint detection model and the 3D freeform deformation techniques to further enhance the quality of the avatar. Another area of improvement is the dependency on the template mesh. We used an initial template mesh as input in this study, which shortened the process but showed that it could only work with a template mesh. In future research, we intend to devise a method to automatically generate meshes from scratch based on image-based information extraction elements by extending them to depth maps.

## Acknowledgement

This research was supported by the Chung-Ang University research scholarship grants in 2023 and Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency(KOCCA) grant funded by the Ministry of Culture, Sports and Tourism(MCST) in 2023(Project Name: Development of digital abusing detection and management technology for a safe Metaverse service, Project Number: RS-2023-00227686, Contribution

Rate: 100%)

## References


[1] Baek, S.-Y., & Lee, K., "Parametric human body shape modeling framework for human-centered product design," *Computer-Aided Design*, 44(1), 56–67, 2012. Article (CrossRef Link)

[2] Bogo, F., Romero, J., Loper, M., & Black, M. J., "Faust: Dataset and evaluation for 3d mesh registration," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 3794–3801, 2014. Article (CrossRef Link)

[3] Buchanan, P., Mukundan, R., & Doggett, M., "Automatic single-view character model reconstruction," in *Proc. of the international symposium on sketch-based interfaces and modeling*, 5–14, 2013. Article (CrossRef Link)

[4] Chen, S., Zhang, K., Shi, Y., Wang, H., Zhu, Y., Song, G., An, S., Krist-jansson, J., Yang, X., & Zwicker, M., "Panic-3d: Stylized single-view 3d reconstruction from portraits of anime characters," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21068–21077, 2023. Article (CrossRef Link)

[5] Duan, H., Lin, K.-Y., Jin, S., Liu, W., Qian, C., & Ouyang, W., "Trb: A novel triplet representation for understanding 2d human body," in *Proc. of the IEEE/CVF international conference on computer vision*, 9479–9488, 2019. Article (CrossRef Link)

[6] Dvorožák, M., S `ykora, D., Curtis, C., Curless, B., Sorkine-Hornung, O., & Salesin, D., "Monster mash: A single-view approach to casual 3d modeling and animation," *ACM Transactions on Graphics (ToG)*, 39(6), 1–12, 2020. Article (CrossRef Link)

[7] Han, X., Gao, C., & Yu, Y., "Deepsketch2face: A deep learning based sketching system for 3d face and caricature modeling," *ACM Transactions on graphics (TOG)*, 36(4), 1–12, 2017. Article (CrossRef Link)

[8] He, K., Zhang, X., Ren, S., & Sun, J., "Deep residual learning for image recognition," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 770–778, 2016. Article (CrossRef Link)

[9] He, T., Xu, Y., Saito, S., Soatto, S., & Tung, T., "Arch++: Animation-ready clothed human reconstruction revisited," in *Proc. of the IEEE/CVF international conference on computer vision*, 11026–11036, 2021. Article (CrossRef Link)

[10] Koestinger, M., Wohlhart, P., Roth, P. M., & Bischof, H., "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. of IEEE international conference on computer vision workshops (ICCV workshops)*, 2144–2151, 2011. Article (CrossRef Link)

[11] Lee, W., Gu, J., & Magnenat-Thalmann, N., "Generating animatable 3d virtual humans from photographs," *Computer Graphics Forum*, 19(3), 1–10, 2000. Article (CrossRef Link)

[12] Li, S., Ke, L., Pratama, K., Tai, Y.-W., Tang, C. -K., & Cheng, K.-T., "Cascaded deep monocular 3d human pose estimation with evolutionary training data," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 6172–6182, 2020. Article (CrossRef Link)

[13] Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J., "Learning a model of facial shape and expression from 4d scans," *ACM Trans. Graph.*, 36(6), 1-17, 2017. Article (CrossRef Link)

[14] Liu, J. -J., Hou, Q., Cheng, M.-M., Wang, C., & Feng, J., "Improving convolutional networks with self-calibrated convolutions," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 10096–1010, 2020. Article (CrossRef Link)

[15] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J., "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, 34(6), 1–16, 2015. Article (CrossRef Link)

[16] Lun, Z., Gadelha, M., Kalogerakis, E., Maji, S., & Wang, R., "3d shape reconstruction from sketches via multi-view convolutional networks," in *Proc. of 2017 International Conference on 3D Vision (3DV)*, 67–77, 2017. Article (CrossRef Link)

[17] Ma, Y. -Y., Zhang, H., & Jiang, S.-W., "Realistic modeling and animation of human body based on scanned data," *Journal of Computer Science and Technology*, 19(4), 529-537, 2004. Article (CrossRef Link)
[18] Magnenat-Thalmann, N., Seo, H., & Cordier, F., "Automatic modeling of virtual humans and body clothing," *Journal of Computer Science and Technology*, 19, 575–584, 2004. Article (CrossRef Link)
[19] Saito, S., Simon, T., Saragih, J., & Joo, H., "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 81-90, 2020. Article (CrossRef Link)
[20] Sederberg, T. W., & Parry, S. R., "Free-form deformation of solid geometric models," *ACM SIGGRAPH Computer Graphicsin*, vol. 20, no. 4, 151–160, 1986. Article (CrossRef Link)
[21] Seo, H., & Magnenat-Thalmann, N., "An automatic modeling of human bodies from sizing parameters," in *Proc. of the 2003 symposium on Interactive 3D graphics*, 19–26, 2003. Article (CrossRef Link)
[22] Sun, K., Xiao, B., Liu, D., & Wang, J., "Deep high-resolution representation learning for human pose estimation," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 5686-5696, 2019. Article (CrossRef Link)
[23] Taylor, A., & Unver, E., "An experimental study to test a 3d laser scanner for body measurement and 3d virtual garment design in fashion education," in *University of Wales*, pp. 1-14, 2005. Article (CrossRef Link)
[24] Tneb, R., Seidl, A., Hansen, G., & Pruett, C., "3-d body scanning-systems, methods and applications for automatic interpretation of 3d surface anthropometrical data," in *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, 44(38), 844–847, 2000. Article (CrossRef Link)
[25] Wang, C. C., Wang, Y., Chang, T. K., & Yuen, M. M., "Virtual human modeling from photographs for garment industry," *Computer-Aided Design*, 35(6), 577–589, 2003. Article (CrossRef Link)
[26] Xiao, Y., & Siebert, J., "Building superquadric men from 3-d whole-body scan data," in *Proc. of 4th IEEE Chapter Conference on Applied Cybernetics*, 82–88, 2005. [Online]. Available: http://eprints.gla.ac.uk/91023/
[27] Xiu, Y., Yang, J., Cao, X., Tzionas, D., & Black, M. J., "Econ: Explicit clothed humans obtained from normals," *arXiv preprint arXiv:2212.07422*, 2022. Article (CrossRef Link)
[28] Xiu, Y., Yang, J., Tzionas, D., & Black, M. J., "Icon: Implicit clothed humans obtained from normals," in *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13286–13296, 2022. Article (CrossRef Link)
[29] Zheng, Z., Yu, T., Liu, Y., & Dai, Q., "Pamir: Parametric model-conditioned implicit representation for image-based humanreconstruction," *IEEE transactions on pattern analysis and machine intelligence*, 44(6), 3170–3184, 2022. Article (CrossRef Link)
[30] Qiu, Yuda, et al., "3dcaricshop: A dataset and a baseline method for single-view 3d caricature face reconstruction," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. Article (CrossRef Link)
[31] Wu, Qianyi, et al., "Alive caricature from 2d to 3d," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. Article (CrossRef Link)
[32] S. Park, M. Ji and J. Chun, "2D Human Pose Estimation based on Object Detection using RGB-D information," *KSII Transactions on Internet and Information Systems*, vol. 12, no. 2, pp. 800-816, 2018. Article (CrossRef Link)
[33] N. Ratyal, I. Taj, U. Bajwa and M. Sajid, "Pose and Expression Invariant Alignment based Multi-View 3D Face Recognition," *KSII Transactions on Internet and Information Systems*, vol. 12, no. 10, pp. 4903-4929, 2018. Article (CrossRef Link)
[34] Pinterest. [Online]. Available: https://www.pinterest.co.kr/

**SANGHYUN BYUN** received his Bachelor's degree in Multimedia from Hannam University in 2022. He is currently pursuing a Master's degree at The Graduate School of Advanced Imaging Science, Multimedia & Film at Chung-Ang University, specializing in Entertainment Technology. His research is centered around artificial intelligence and computer vision, with a particular focus on Human Pose Estimation and 3D Avatar Generation.

**BUMSOO KIM** received the B.S degree in Art and Technology from Chung-Ang University in 2023, South Korea. He is currently an AI Researcher with the VIVE STUDIOS, South Korea. His research interests include style transfer, stylization(cartoonization) and video-level face re-aging, face swap.

**YONGHOON JUNG** received his Bachelor's degree in Computer Engineering from Sungkyul University in 2022. He is currently pursuing a Master's degree at The Graduate School of Advanced Imaging Science, Multimedia & Film at Chung-Ang University, specializing in Entertainment Technology. His research is centered around artificial intelligence and computer vision, with a particular focus on synthetic data generation and domain adaptation techniques. He aims to apply these cutting-edge technologies to solve complex issues in the real world. His dedication to his field is evident as he continues to explore innovative solutions to enhance technological applications.

**WONSEOP SHIN** obtained a BS degree in Computer Science from Sungkyul University in South Korea from 2017 to 2022. He is currently earning a MS degree from Chung-Ang University, Graduate School of Advanced Imaging Science in South Korea since 2023. His areas of interest include Deep Learning, Object Detection, Computer Vision, as well as Virtual and Augmented Reality.

**SANGHYUN SEO** received the B.S. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 1998, and the M.S. and Ph.D. degrees from the GSAIM Department, Chung-Ang University, in 2000 and 2010, respectively. He was a Senior Researcher with G-Inno Systems, from 2002 to 2005. He was a Postdoctoral Researcher with Chung-Ang University, in 2010, and the LIRIS Laboratory, Lyon 1 University, from February 2011 to February 2013. He has worked at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, from May 2013 to February 2016. He has also worked at Sungkyul University, from March 2016 to February 2019. He is currently a Faculty Member with the College of Art and Technology, Chung-Ang University. His research interests include computer graphics, non-photorealistic rendering and animation, real-time rendering using GPU, VR/AR, and game technology. He has been a program committee member of many international conferences and workshops. He has been a Reviewer of Multimedia Tools and Applications (MTAP), Computers and Graphics (Elsevier), U.K., the Journal of Supercomputing (JOS), and The Visual Computer (Springer). He has edited a number of international journal special issues as a Guest Editor, such as the Journal of Real-Time Image Processing, the Journal of Internet Technology, and Multimedia Tools and Applications. He has been an Associate Editor of the Journal of Real-Time Image Processing, since 2017.