Generative AI-driven Graphic Pipeline for Web-based Editing of 4D Volumetric Data

Ye-Won Jang[1], Jung-Woo Kim[2], Hak-Bum Lee[2]
and Young-Ho Seo[2,*]

[1]*Omotion Inc., Korea*
[2]*Kwangwoon University, Korea*
E-mail: yhseo@kw.ac.kr; yw.jang@omotion.co.kr
[*]*Corresponding Author*

Received 20 October 2024; Accepted 03 December 2024

Abstract

This paper proposes a novel approach to adding and editing clothing and movement of 4D volumetric video data in a web-based environment. While significant advancements have been made in 3D modeling and animation, efficiently editing 3D mesh data produced in sequence remains a challenging problem. Since 3D mesh data synthesized from multiple cameras exists continuously over time, modifying a single 3D mesh model requires consistent editing across multiple frames. Most existing methods focus on single meshes or static 3D models, limiting their ability to handle the complexity of time-varying 3D mesh sequences. The method proposed in this paper targets 3D volumetric sequences synthesized from multiple cameras. It utilizes deep learning networks to estimate body poses, facial features, and hand shapes from RGB images, generating 3D models using the SMPL-X method. Subsequently, an algorithm is applied to segment the 3D mesh, separating and combining the head and torso of the model to create a new 3D model. In the web-based environment, this process makes the data editable, allowing for adding new motions or replacing clothing, which can be seamlessly

Journal of Web Engineering, Vol. 24_1, 135–162.
doi: 10.13052/jwe1540-9589.2416

composited into the existing sequence video. The proposed method enables editing and modification of various types of 3D mesh sequences, facilitating enhancements to existing sequences, such as changing the motion of characters or replacing their clothing, thereby improving the overall quality of 3D content creation in online applications.

Keywords: Web-based services, 4D volumetric, 3D model, SMPL-X, virtual human, online editing.

1 Introduction

With the recent growth in the virtual reality (VR), augmented reality (AR), and mixed reality (MR) industries, the demand for high-quality 3D content that seamlessly connects the real and virtual worlds has been on the rise [1]. To achieve a natural integration between these two realms, virtual content must provide a level of realism comparable to the real world, necessitating 3D models that accurately reflect real-world subject appearance and movements. Consequently, volumetric content, which captures subjects in 3D data form and enables lifelike representation of appearance and motion, is being increasingly utilized across various fields.

Sequences or videos of 3D models are representative examples of volumetric content created by combining sequential frames of 3D volumetric data. This 3D volumetric data is generated by continuously capturing a subject using multiple cameras over time and then synchronizing the captured footage in the same temporal order to form a single 3D mesh model [2, 3]. While volumetric data offers the advantage of faithfully recording a subject's appearance and movement, it is not easy to edit since it is produced as a continuous sequence over time [4].

Due to the temporal continuity of the 3D mesh data, modifying one 3D mesh model requires editing all the 3D mesh models across multiple frames. Typically, 3D mesh data synthesized through photogrammetry using footage from various angles will have different mesh structures and topologies. Although a single object might appear to have the same shape across frames, the mesh topology differs from frame to frame. Therefore, consistently editing temporally variable 3D meshes is time-consuming and costly [1, 4]. Consequently, an efficient method for modifying the motion of volumetric meshes is essential to overcome these challenges.

Existing research on volumetric model animation has primarily focused on the SMPL/SMPL-X models. The SMPL model is widely used for

reconstructing human 3D shapes and poses from images, and its simple mesh structure allows for the efficient representation of various human postures and forms [5]. The SMPL-X model extends the capabilities of SMPL by enabling more detailed expressions, including facial features and hand movements [6]. Most papers in this field focus on generating 3D human avatars by estimating the parameters of SMPL/SMPL-X models from a single image or video [6–13]. However, since SMPL-based studies struggle to capture detailed facial expressions and finger movements accurately, there has been a growing shift toward using the SMPL-X model in recent research.

This paper focuses on capturing the full body pose and shape, including hands and facial expressions, to achieve more accurate 3D modeling [6, 10–15]. However, these methods often fail to capture elements separate from the body, such as hair, leading to discrepancies between the original volumetric model and the reconstructed 3D mesh, making it challenging to reproduce the subject's appearance faithfully. Additionally, many studies rely on the facial expression parameters estimated by the SMPL-X model, which limits the accurate representation of complex expressions or subtle facial movements. To address these issues, this paper proposes a method that, while using the SMPL-X model for pose estimation, directly incorporates the original volumetric data for facial regions. This approach allows for more efficient editing and refinement of the synthesized 3D mesh data, particularly in maintaining fidelity to the original facial features of the volumetric model.

The method proposed in this paper consists of three stages. In the first stage, a deep learning network processes the 3D volumetric data synthesized from multiple cameras. This captures the basic body shape, facial expressions, and finger movements in SMPL-X format, generating a complete 3D model. Using a mesh segmentation algorithm, the model's head and torso are separated in the second stage. This begins by generating a 3D skeleton using OpenPose, and the initial mesh segmentation is performed based on the direction vectors of the generated skeleton. After this, manual refinement is carried out further to enhance the separation of the head and torso. This process involves calculating the vector from the neck joint to the head joint, enabling precise separation of the head region. Finally, the separated volumetric head and SMPL-X torso are combined to create a new 3D model in the third stage. To add new motion animations, the SMPL-X skeleton is applied to the newly created 3D model through a rigging process, resulting in a fully animated and functional 3D model.

This paper is structured as follows: Section 2 explains creating 3D volumetric models. Section 3 introduces the algorithm for editing volumetric

videos. Section 4 presents the results obtained using the proposed algorithm, and Section 5 concludes the paper with a summary and final remarks.

2 4D Volumetric Content

This chapter provides a detailed explanation of the process through which a 3D volumetric model is produced. Volumetric capture is divided into three stages: capturing, volumetric synthesis, and post-processing. It explains capturing, which involves recording a subject in a volumetric studio, and the reconstruction stage, where the captured footage is synthesized and converted into 3D mesh data. Lastly, it discusses the editing stage, where the synthesized and generated volumetric model is modified and corrected.

2.1 Capturing

The first stage of volumetric capture is the setup of the studio environment. Essential factors in this stage include rigging and lighting configuration, which require careful consideration to precisely capture the subject's 3D information. First, the rig setup begins with cameras that can cover a 360-degree view. The cameras are arranged to collect data from all angles of the subject and, in this process, it is crucial to determine the optimal shooting range and resolution, taking into account the size and movement of the subject [16]. This is because volumetric capture requires different depth information from each angle. In this process, adjusting the camera positions significantly impacts the quality of the captured data, so the field of view and overlap of each camera are adjusted.

Next, the lighting setup mainly uses soft light to emphasize the texture and depth information of the subject. Soft light provides uniform illumination to the subject, allowing precise capture of surface details while minimizing unnecessary shadow formation, which keeps the data clean during post-processing. In particular, in volumetric capture, since texture affects the realism of the 3D model, lighting design plays a vital role [17]. Chroma key shooting is also a critical part of volumetric capture. Since the chroma-essential background separates the subject during post-processing, avoiding clothing or props that match the background color is crucial. Appropriate costume and prop selection enhances the accuracy of chroma keying, which directly affects the quality of the result. For this reason, the design and color of the subject must be chosen considering the contrast with the chroma key screen.

Once the studio environment is prepared, a real-time preview system checks the position between the subject and the cameras. At this stage, adjustments are made to the subject's position and angle, as well as the camera's focus and composition, to improve the accuracy of the capture. Additionally, the cameras are synchronized using light sync technology, which is essential for accurately combining camera data in the subsequent 3D modeling phase.

2.2 Reconstruction

Synthesizing a 3D model is based on point clouds generated through photogrammetry or structure from motion (SfM) techniques, which are crucial in precisely digitizing real-world objects. This process typically involves reconstructing 3D shapes from multiple 2D images by matching feature points across images and estimating 3D coordinates from these matches. In the first stage, multiple captured images are analyzed to extract feature points from each image, and these points are matched to establish correlations between the images. This step is usually carried out using algorithms such as SIFT (scale-invariant feature transform) or SURF (speeded-up robust features), which analyze how the same point on an object appears across different images [18]. Based on this feature point matching, the position and orientation of the cameras, along with the 3D coordinates of the object's surface, are estimated using triangulation techniques [19].

The second stage involves constructing a polygon mesh from the generated point cloud. While the point cloud represents the object's surface in 3D space, the points lack connectivity, so a process is required to build a polygonal structure. Typically, techniques such as Delaunay triangulation or Poisson surface reconstruction are employed to define the relationships

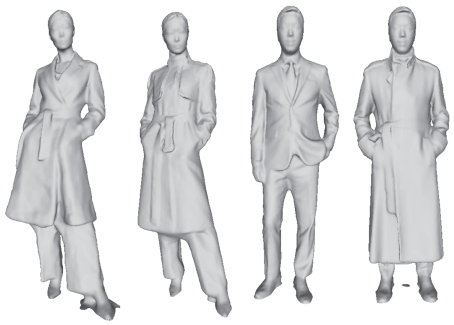


Figure 1 Example of 4D volumetric data.

between points in the cloud and form a mesh by connecting the points into triangles [20].

In the final stage, texture mapping is applied to the generated 3D mesh to enhance realism. Texture mapping involves projecting color and surface detail information obtained from 2D images onto each polygon of the 3D model. This process significantly increases the visual detail of the model, making it resemble the appearance of the actual object. Additionally, lighting and shading effects are applied further to emphasize the texture and depth of the object, ultimately resulting in a more realistic rendered outcome. These steps are essential for achieving high-quality, lifelike 3D representations [21].

2.3 Post-Processing

The process of post-editing a 3D photorealistic model synthesized from multi-view cameras is a crucial procedure aimed at enhancing the quality of the 3D data and maximizing its visual completeness. This process typically involves noise removal, identification and deletion of erroneous point clouds, mesh reconstruction, texture map adjustment, color correction, and the addition of lighting effects. One of the most common issues in a 3D model is that noise and erroneous points in the point cloud can compromise the model's accuracy. Such noise can arise from various factors, especially in multi-view camera systems, where issues such as insufficient visual overlap between cameras, sensor errors, or environmental factors (e.g., light reflection) may result in incorrect point cloud generation. The noise removal process filters out these erroneous data, improving the quality of the point cloud by identifying and deleting unnecessary or faulty points. This is generally achieved through algorithms like outlier removal or techniques such as statistical noise filtering [22]. After noise removal, mesh reconstruction follows. At this stage, a new 3D mesh is generated based on the cleaned point cloud, improving the smoothness of the model's surface. The goal of mesh reconstruction is to enhance the structural completeness of the model by re-establishing the connections between point clouds, thereby creating a natural and smooth surface. Standard algorithms used here include Poisson surface reconstruction and Delaunay triangulation [20].

The texture map, which plays a critical role in determining the visual appearance of the 3D model, involves applying image data to the model's surface. The model's visual consistency can be compromised if the texture map is incorrectly applied or distorted. Therefore, precise adjustment of the texture map to ensure accurate alignment with the mesh is essential. UV

mapping ensures that each texture is evenly distributed across the model's surface during this process. For photorealistic models, in particular, the detailed refinement of textures significantly contributes to their realism. Lastly, the visual completeness of the model is further enhanced through color correction and the addition of lighting effects. Color correction addresses any color distortions that may have occurred during capture, while lighting effects are added to reinforce the realism of the subject. Techniques such as gamma correction or white balance adjustment are typically employed in color correction to ensure that the model's colors are rendered naturally and consistently. Additionally, lighting effects are applied to give the model a sense of depth and to strengthen its three-dimensional representation.

3 Proposed Pipeline

3.1 Algorithm Overview

This section provides a detailed explanation of the algorithm for editing 4D volumetric content. The process of editing and modifying 3D mesh data produced as a 3D model sequence is divided into four main stages. In the first stage, SMPL-X is utilized to infer the 3D model. The second stage involves segmenting the 3D model (both volumetric and SMPL-X) into two parts, the head, and the torso, allowing for independent processing of each section. In the third stage, the separated head and torso are recombined to create a new 3D model, enabling modifications in motion and attire. Finally, in the fourth stage, these 3D models are linked to produce 4D volumetric content. The proposed algorithm is illustrated in Figure 2.

Each step of the process will be explained in detail. In the first step, the inference of 3D mesh data is performed using the SMPL-X model. SMPL-X is an extended version of the original SMPL model, incorporating more detailed representations of the human body, such as the face, fingers, and toes, which enhances the accuracy of human body modeling and inference [6]. SMPL-X demonstrates improved performance compared to SMPL, particularly in estimating natural human poses and gestures. Moreover, it is structured to effectively handle the complexity of multidimensional data, making it well-suited for detailed human body modeling tasks. In the second stage, the inferred 3D model is segmented into the head and the torso, allowing for independent processing of each part. This is part of a modular approach, which enables fine-tuned manipulation by allowing each body part to be edited separately. Such an approach is beneficial in volumetric editing,

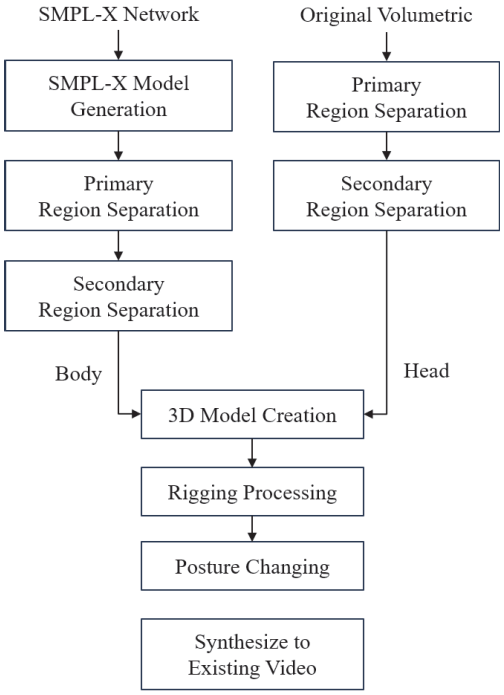


Figure 2 Proposed algorithm.

where the independent handling of complex objects is required. This method has been frequently highlighted in related research as an effective technique for splitting and recombining models. In the third stage, the previously separated head and torso are recombined to generate a new form of the 3D model, serving as the foundation for motion and clothing transformations. Motion transformation is carried out using a motion retargeting algorithm, which naturally applies the movements of one individual to another model [23]. A virtual fitting technique is applied for clothing transformation, designed to automatically adjust the garment based on the model's body shape and motion [24]. The final stage involves sequentially connecting the generated 3D models along the time axis to create the 4D volumetric content. This results in 4D data with spatiotemporal consistency, which can be utilized in real-time within virtual reality (VR) or augmented reality (AR) environments. Temporal volumetric synthesis is critical, ensuring smooth transitions between frames while maintaining continuity [25]. An example in Figure 3 illustrates a detailed algorithm implementation.

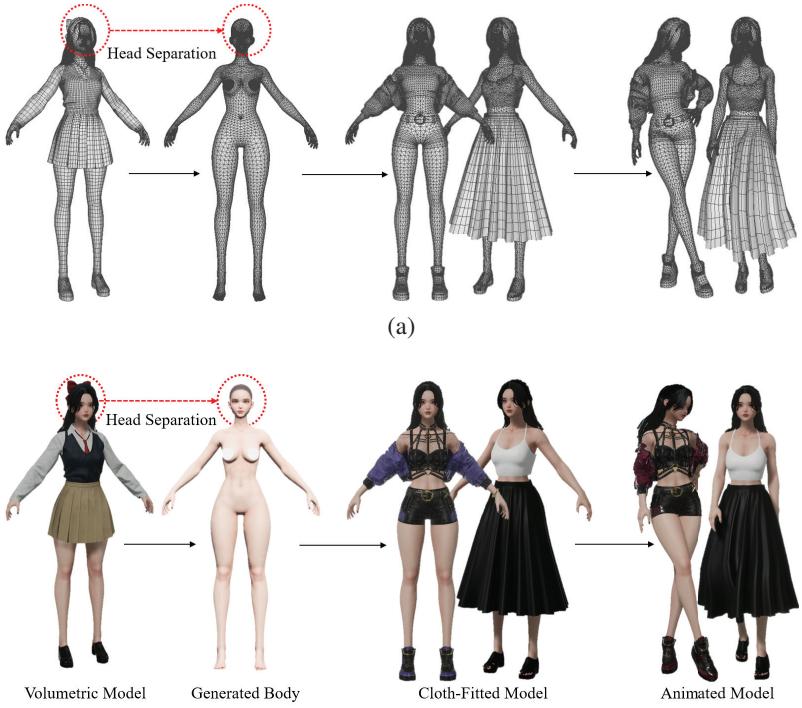


Figure 3 Visual example of the proposed algorithm. (a) 3D model's mesh wireframe, (b) 3D model with applied textures.

3.2 3D Model Inference

Direct modification of volumetric data requires complex computations in high-dimensional space, making it essential to employ deep learning-based models for estimation and refinement. In particular, volumetric data must maintain consistency across time and space, necessitating algorithms capable of precise data processing. To address this, this paper proposes a methodology that uses deep learning models to estimate body shape and motion from volumetric data, then editing and refining 3D mesh data based on these estimations.

A widely used model in this context is the skinned multi-person linear (SMPL) model, a linear model for representing 3D human body meshes extensively applied in both AI and graphics fields [5]. The SMPL model parameterizes human body shape and pose, allowing the generation of a

3D body model from a single 2D image. In this process, shape and pose parameters are used to model a body shape similar to the person in the input image. However, the SMPL model has limitations in expressing detailed regions such as hands and faces. The SMPL-X model, an extension of SMPL, overcomes this by incorporating facial expressions and hand movements, enabling more detailed human body modeling [6]. SMPL-X offers better performance, particularly in tasks where facial expressions and hand gestures are critical, making it an essential tool in applications like virtual reality (VR) and augmented reality (AR), where human interaction plays a central role.

In this paper, we employ the ExPose (expressive pose and shape regression) model, an advancement of the SMPL-X model, to estimate and modify body shape and pose in volumetric data. ExPose can accurately estimate the pose and shape of the body, face, and hands from RGB images, and it generates 3D mesh models that closely resemble the volumetric data [11].

ExPose operates based on the SMPL-X framework, directly regressing complex 3D data to swiftly and accurately estimate detailed human body parameters. These key parameters are broadly categorized into three types. First, the shape parameter represents the overall body shape, defining global features such as the length and size of the body. Second, the expression parameter is used to model facial expressions, allowing for detailed reflection of emotional expressions on the face. Lastly, the pose parameter defines the rotation of body joints using an axis-angle representation, capturing the body's posture and movement [11]. The 3D objects generated by ExPose in SMPL-X format reflect the three-dimensional characteristics of the volumetric data, producing a temporally consistent mesh that closely mirrors the body shape and posture of the original volumetric data. Errors in this process mainly arise from inferring 3D information from 2D images. Since 2D images inherently lack 3D depth information, errors during the 3D reconstruction process are inevitable [9].

The SMPL-X model is designed to effectively infer human appearance from 2D images, yet it has limitations in capturing all fine-grained physical characteristics with high precision. As a result, it may not accurately represent minute facial wrinkles, subtle expressions, or the full range of body shapes with intricate detail. The quality of 4D volumetric data can vary depending on the shooting environment and synthesis conditions. The synthesis quality of the volumetric model we use aligns well with the capabilities of SMPL-X, as previously described. Therefore, employing SMPL-X to represent the volumetric model is highly suitable for this paper.

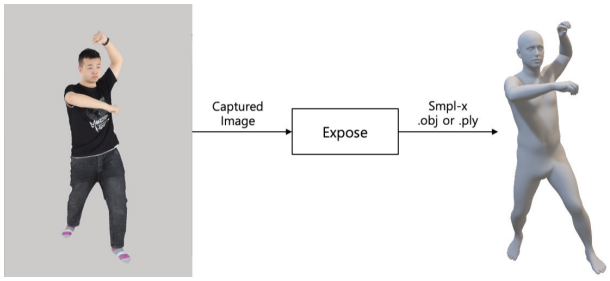


Figure 4 Definition of input (2D image) and output (SMPL-format 3D mesh) of the ExPose deep learning model.

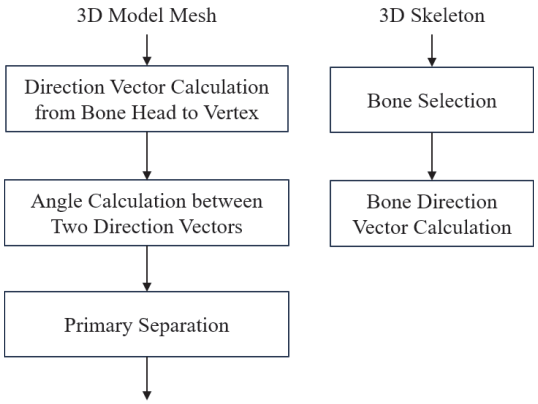


Figure 5 The first process for separating the head from the torso.

3.3 Separation of 3D Model

This is the process of separating the head from the original volumetric data and the torso from the SMPL-X model to create a new 3D model. The separation process is divided into two stages. The first stage utilizes deep learning to segment the head and torso regions. In contrast, the second stage involves manual refinement to achieve a more precise separation based on the initial segmentation results.

OpenPose is first utilized to generate 3D skeletons for the volumetric and SMPL-X models to perform the initial region segmentation. In this process, projection images are created by viewing the 3D mesh from four directions—front, back, left, and right—to estimate the 3D pose of the mesh. Subsequently, the OpenPose library is employed to extract 2D joint positions from the projection images, and the 3D joint positions are calculated by

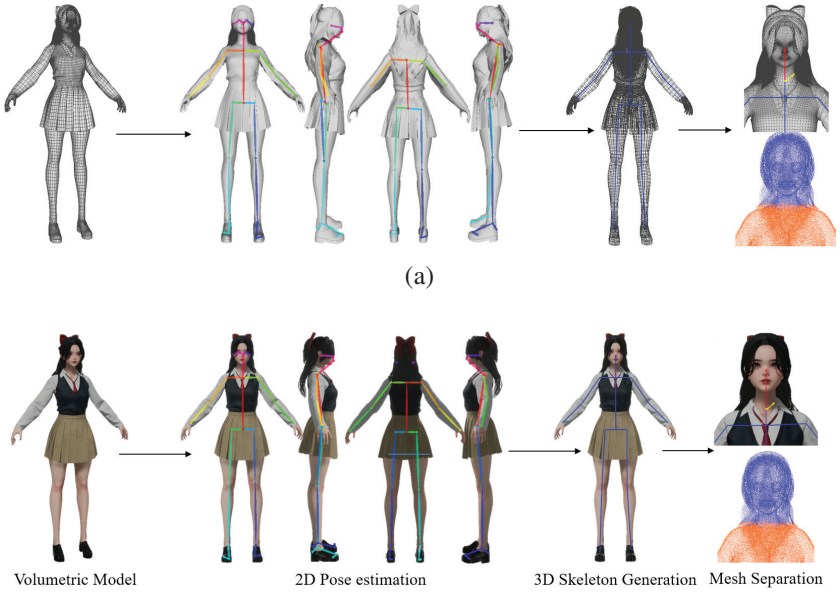


Figure 6 Visual example of the first-stage separation algorithm (a) separation process illustrated with the 3D model's mesh wireframe, (b) separation process shown with the textured 3D model.

determining the intersections in 3D space. Based on the joint information generated in this way, the 3D mesh regions corresponding to the head and torso of both the volumetric and SMPL-X models are separated.

The segmentation of mesh regions is carried out by calculating the direction vectors of the skeleton, which are derived from the 3D coordinates of the two joints forming each skeletal segment. To separate the head from the torso, the vector directed from the neck joint to the head joint is defined as the skeleton's direction vector. At this point, if the vector from the starting point of the skeleton to a vertex on the mesh forms an angle within 90 degrees of the skeleton's direction vector, that vertex is considered part of the bone and can be used to separate the head region. This process is illustrated in Figure 6.

In the second stage of region segmentation, manual refinements correct the errors in the initial segmentation results, ultimately producing a properly separated mesh. While the first-stage segmentation is performed using automated algorithms, various errors may arise. One common issue is the frequent

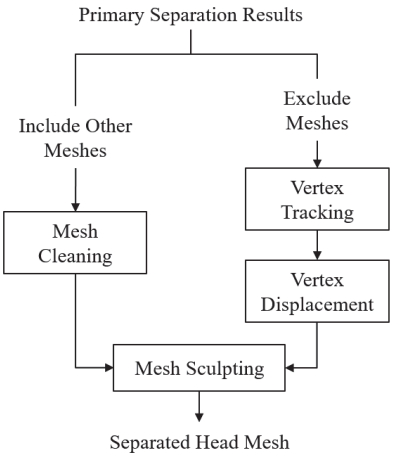


Figure 7 Algorithm for the manual process.

mixing of the head mesh with other body parts. The key focus of the second-stage segmentation is to meticulously correct these errors, ensuring that the final mesh segmentation is accurate and precise.

First, if the head mesh in the initial segmentation results contains portions of other regions, errors are corrected based on the vertex and face information of the affected areas. Vertices and faces from incorrectly included regions are removed using a mesh cleaning technique, identifying and deleting unnecessary parts according to specific criteria [26]. This process must remove unwanted parts while maintaining the mesh's connectivity.

Second, if parts of the head region are missing or distorted during the segmentation process, a correction is made by tracking the vertices of the missing areas. Specifically, the nearest neighbor technique is employed to locate the position and coordinates of vertices closest to the excluded areas [27]. By referencing the coordinates of these vertices, the distorted sections are moved on the mesh, thereby correcting the shape of the head mesh. This ensures the continuity and morphological consistency of the mesh and is based on vertex displacement techniques, which adjust the mesh's shape by utilizing the distances and coordinates between adjacent vertices [28].

This approach to correcting mesh distortions is particularly useful for complex regions like the head. It is done through a hybrid method combining manual work with automated algorithms. A finely segmented head mesh is ultimately obtained by precisely correcting the errors that occurred during

(a) (b) (c)

Figure 8 Visual example of the second-stage separation algorithm. (a) Volumetric head, (b) refinement mesh, (c) final head.

the initial segmentation, as shown in Figure 8. Figure 8 provides an example of a 3D mesh of the separated face.

3.4 Synthesis and Conversion

This produces a 4D volumetric video by integrating edited frames and sequences of a new 3D model into the original volumetric sequence. In this process, synchronization is performed to connect frames and sequences, where motion is added seamlessly or costumes are replaced, to the flow of the existing volumetric data sequence. Additionally, post-processing is carried out to apply gradual changes, ensuring smooth transitions between the original sequence and the modified frames. Through this process, a consistent and natural 4D volumetric video is completed. An example of the overall process is shown in Figure 9.

The process involves integrating the edited frames and sequences of the new 3D model into the original volumetric sequence to produce a 4D volumetric video. This step includes synchronization, ensuring that the newly added frames or sequences—such as those with modified movements or clothing—are seamlessly connected to the existing volumetric data sequence flow. Additionally, post-processing is applied to create smooth transitions between the original sequence and the modified frames, using gradual adjustments to maintain fluidity. Through this process, a consistent and natural 4D volumetric video is achieved. An example of the overall procedure is illustrated in Figure 10. Cloth simulation is employed to apply clothing to the SMPL-X model. 4D volumetric sequences are used to preserve natural forms

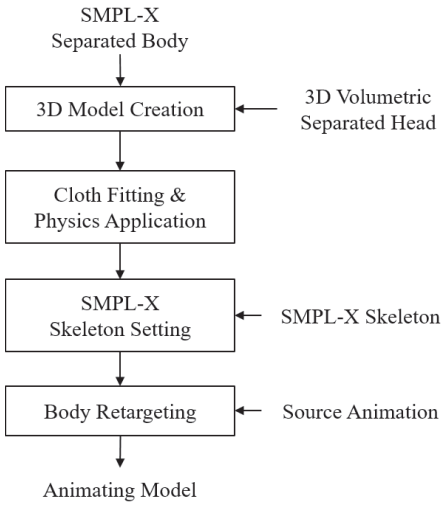


Figure 9 Production and editing process.

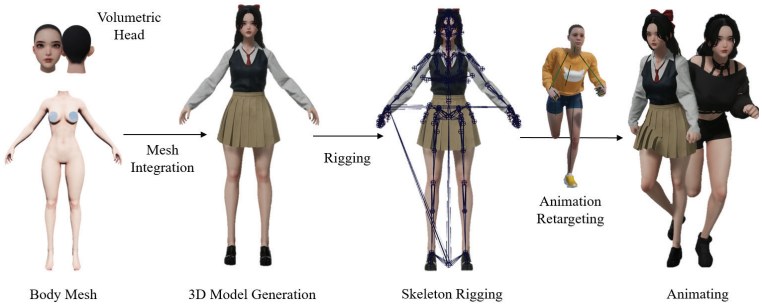


Figure 10 Visual example of the volumetric video integration process.

and movements. Therefore, to better reflect the characteristics of 4D volumetric data, we use physics-based cloth simulation rather than a rigging-based approach.

Detail-preserving algorithms are essential to minimize distortion when applying textures to complex surfaces. Utilizing level of detail (LOD) and deformation correction algorithms allows natural textures to be applied to 3D model surfaces. Additionally, AI-based upscaling and GAN models can convert low-resolution textures to high-resolution and correct distortions. To maintain temporal continuity in volumetric data, spatiotemporal interpolation techniques are introduced to ensure that changes between keyframes and temporal shifts in texture maps are smoothly connected. This approach

Table 1 Experimental environment and conditions

Category	Settings
Camera	SONY Full-frame mirror less
Number of units	40 units (360-degree full coverage)
Camera rig	Movable rig, 2-tier structure
Frame settings	Still shooting 60FPS, video shooting 120FPS
Synchronization	Light sync
Calibration	Calibration board

Figure 11 Shooting environment for making the 4D volumetric model.

helps maintain graphic consistency even in moving objects or dynamically changing environments.

4 Implementation Result

4.1 Environment

This study constructed a studio and capture system using 40 full-frame cameras, as depicted in Figure 11. Multiple cameras are arranged within the rig to provide 360-degree coverage from all directions. The rig's position is adjusted to set a capture range between 8 and 12 m, depending on the subject and its movement.

Figure 12 presents a 3D volumetric sequence generated through a 4D volumetric content production process. In this process, multiple cameras were used to capture the subject from various angles, and the captured video data were subsequently merged to create a 3D model. The resulting 3D models were then arranged sequentially to form the complete sequence. Figures 12(a) and 12(b) depict specific frames from this sequence. The model texture was applied at 8K resolution, with the mesh in Figure 12(a) consisting of 48,274

(a) (b)

Figure 12 Frames from the original 4D volumetric content. (a) Model 1, (b) model 2.

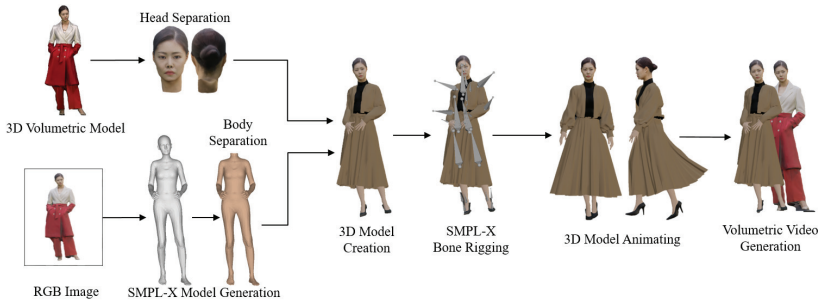


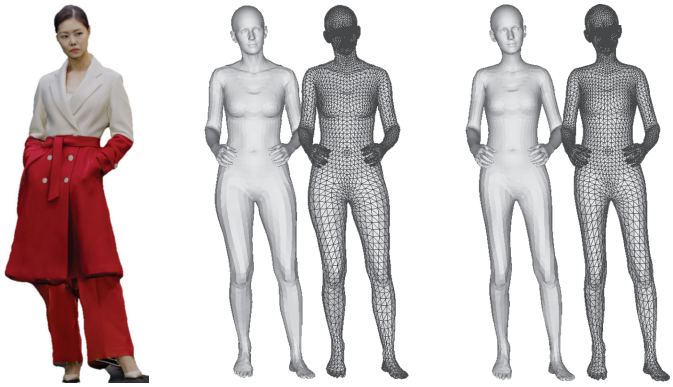
Figure 13 Visual example of the proposed algorithm.

faces and the mesh in Figure 12(b) consisting of 50,813 faces. The number of faces in the mesh varies across frames due to differing compositing conditions for each frame in the 4D volumetric data.

Figure 13 illustrates the results obtained through each step of the algorithm proposed in this paper. First, an SMPL-X model was generated from the 2D images of the 3D volumetric model, and a mesh separation algorithm was employed to isolate the head of the 3D volumetric model and the torso of the SMPL-X model. Subsequently, the separated meshes were merged, and a new outfit was applied to create a modified 3D model. Finally, rigging and animation were added to the 3D model, resulting in the adjusted motions of the original volumetric model.

4.2 Inference Results of the Model using SMPL-X

Figure 14 shows the SMPL-X 3D mesh results inferred from the input volumetric model frames that require modification. Using a deep learning-based



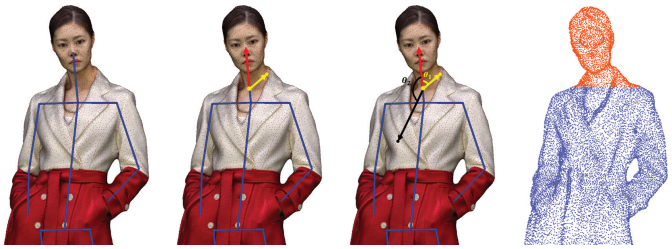
(a) (b) (c)

Figure 14 3D volumetric data and SMPL-X results. (a) Original volumetric data of the input frame, (b) SMPL-X model inferred from the input frame, (c) adjusted mesh of the inferred SMPL-X model.

SMPL-X model, the size, body shape, and motion similar to the volumetric model were modeled from the 2D images. Figure 14(a) displays the volumetric data used as input, while Figure 14(b) illustrates the initial SMPL-X model inferred from this data. The initial model's body shape differs from the volumetric model's, necessitating post-processing to reduce this discrepancy. Figure 14(c) presents the final result, where the body shape has been adjusted through post-processing to more closely resemble the volumetric model in Figure 14(a).

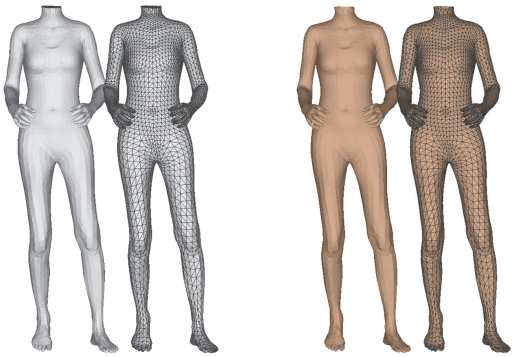
4.3 Segmentation Results of the 3D Model

Figure 15 illustrates the process and outcome of the primary separation algorithm used to divide the head and torso of a 3D model. OpenPose was utilized to generate the model's 3D skeleton, which served as the basis for the separation process. In Figure 15(a), the bone connecting the neck joint to the head joint was selected based on the generated joint data, and the direction vector of this bone was calculated. Figure 15(b) shows the calculation of the direction vectors from the selected bone's one-third point to each mesh vertex. Figure 15(c) depicts calculating the angle between the two direction vectors to distinguish the head from the torso, with separation proceeding based on angles within 90 degrees. Finally, Figure 15(d) presents the result of the separation, where the head and torso meshes were divided based on the calculated angles.



(a) (b) (c) (d)

Figure 15 First-stage separation algorithm result. (a) Bone selection and direction vector calculation, (b) direction vector calculation from the 1/3 point of the bone to the vertex, (c) direction vector angle calculation, (d) mesh region separation.

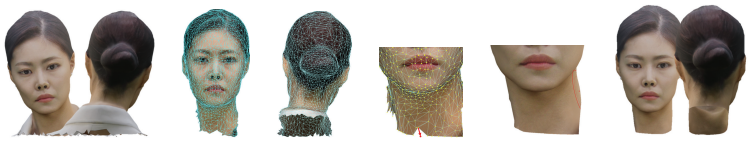


(a) (b)

Figure 16 Results of the primary mesh separation. (a) SMPL-X 3D body model with the face removed, (b) 3D body model with applied texture.

Figure 16 shows the final result of separating the head of the volumetric model and the torso of the SMPL-X model using the primary separation algorithm. Figure 16(a) presents the SMPL-X model after separating the torso. Figure 16(b) illustrates the outcome of applying a texture to the SMPL-X model generated to match the skin tone of the volumetric model.

Figure 17 illustrates the process and outcome of the secondary separation algorithm used to divide the head and torso of the 3D model. This process involved manually refining errors that occurred during the primary separation. Figure 17(a) shows the initial result of the head separation from the volumetric model after the first separation process. Figures 17(b) and 17(c) display the results of applying a mesh cleaning technique to remove unnecessary vertices and faces from the model, with Figure 17(b) showing the



(a) (b) (c) (d) (e) (f)

Figure 17 Second-stage separation algorithm result. (a) First-stage separation result, (b) vertex and face removal (front), (c) vertex and face removal (back), (d) vertex displacement, (e) mesh sculpting, (f) final result of head separation.

(a) (b) (c)

Figure 18 Synthesized 3D model and rigging results. (a) Original volumetric model, (b) result of applying new clothing to the newly generated 3D model, (c) result of rigging with a SMPL-X skeleton.

front view and Figure 17(c) the rearview. Figure 17(d) demonstrates the use of the nearest neighbor technique to locate the positions of adjacent vertices for the missing areas, followed by applying vertex displacement to correct those regions. Figure 17(e) presents the result of finely adjusting the distorted areas on the mesh using a sculpting technique. Finally, Figure 17(f) shows the fully separated mesh, confirming that a clean and refined mesh was achieved through the secondary separation process.

4.4 Results of Synthesis and Transformation of the 3D Model

Figure 18 presents the results after combining the separated head and torso meshes, applying new clothing, and configuring the skeleton using the SMPL-X framework to generate a 3D model. Figure 18(a) shows the original volumetric model data. Figure 18(b) displays the 3D model with newly

Figure 19 Results of implementing new movements on the 3D model. (a) Source motion 1, (b) retargeting of source motion 1 onto the 3D model, (c) source motion 2, (d) retargeting of source motion 2 onto the 3D model.

applied clothing, demonstrating the ability to change the outfit of the original volumetric model. Finally, Figure 18(c) illustrates the result of rigging the generated 3D model with the SMPL-X skeleton, enabling the model to be animated and moved.

Figure 19 shows the result of animating a 3D model using retargeting technology. The motion data of the source character depicted in Figure 19(a) and Figure 19(c) was applied to the target 3D model using an IK retargeting method, and the 3D volumetric model was animated, as shown in Figures 19(b) and 19(d). In this process, the joint positions and angle information of the source character were mapped to the corresponding joints of the target 3D model to achieve natural movement. Despite structural differences between the source and target characters, the source motion was successfully identically applied to the target model.

4.5 4D Volumetric Content Production Results

Figure 20 visually presents the step-by-step results from the creation to the animation of the 3D model. The head of the separated volumetric model was combined with the SMPL-X torso to generate a new 3D volumetric model, and the SMPL-X skeleton was rigged into the model to allow for movement modification. Using a body retargeting method, source animations were applied to the volumetric 3D model, enabling the animation of various movements.

Figures 21 and 22 display the results of applying motion and clothing changes to the original 3D volumetric data arranged as a sequence of frames.

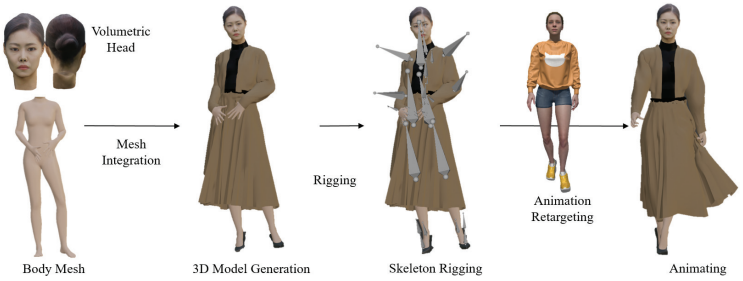


Figure 20 Results of the entire editing process.

Figure 21 Results of motion and clothing change (long skirt) applied to the original 3D volumetric data.

Figure 22 Results of motion and clothing change (short skirt) applied to the original 3D volumetric data.

Various movements, such as walking and rotation, were implemented, and the motion reflected the physical properties of the clothing. These results successfully modify the volumetric model's clothing and movements, with smooth transitions between frames and consistent animation throughout.

5 Conclusion

This paper proposes a novel approach for editing and modifying sequential 3D mesh data captured by multiple cameras in a web-based environment. First, a deep learning network is used to estimate body posture, shape, facial features, and hand details from RGB images, which are then converted into a 3D volumetric model using the SMPL-X framework. Next, an algorithm is employed to segment the 3D mesh, separating the torso of the SMPL-X model from the head of the volumetric model through a two-step process, followed by their combination to create a new 3D model. Clothing, accessories, and footwear are applied to the generated model, and simple motions are added using SMPL-X skeleton rigging and body retargeting techniques, allowing for the replacement or insertion of new 3D volumetric data into existing sequences. This approach enables the creation of edited 3D volumetric mesh sequences and facilitates the addition and modification of movements or outfits in already-produced 3D volumetric sequences. This method is expected to significantly enhance content creation flexibility, with potential applications across various industries such as film, gaming, and AR/VR environments based on online services.

Acknowledgements

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Development of a webtoon IP utilization platform using 3D animating AI capable of accurate Korean expression, Project Number: RS-2024-00397183, Contribution Rate: 100%. This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2022-2020-0-01846) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation)).

References

[1] R. Pandey, A. Tkach, S. Yang, P. Pidlypenskyi, J. Taylor, R. Martin-Brualla, A. Tagliasacchi, G. Papandreou, P. Davidson, C. Keskin *et al.*, "Volumetric capture of humans with a single rgbd camera via semi-parametric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9709–9718.

[2] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu, "Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5746–5756.

[3] J. Starck and A. Hilton, "Surface capture for performance-based animation," *IEEE computer graphics and applications*, vol. 27, no. 3, pp. 21–31, 2007.

[4] M. Moynihan, S. Ruano, A. Smolic *et al.*, "Autonomous tracking for volumetric video sequences," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1660–1669.

[5] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[6] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.

[7] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1954–1963.

[8] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Detailed human avatars from monocular video," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 98–109.

[9] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7122–7131.

[10] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2252–2261.

[11] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 20–40.

[12] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 965–10 974.

[13] Y. Feng, J. Yang, M. Pollefeys, M. J. Black, and T. Bolkart, "Capturing and animation of body and clothing from monocular video," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–9.

[14] X. Zhao, Y.-T. Hu, Z. Ren, and A. G. Schwing, "Occupancy planes for single-view rgb-d human reconstruction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3633–3641.

[15] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, "Collaborative regression of expressive bodies using moderation," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 792–804.

[16] O. Schreer, I. Feldmann, S. Renault, M. Zepp, M. Worchel, P. Eisert, and P. Kauff, "Capture and 3d video processing of volumetric video," in *2019 IEEE International conference on image processing (ICIP)*. IEEE, 2019, pp. 4310–4314.

[17] K. Guo, P. Lincoln, P. Davidson, J. Busch, X. Yu, M. Whalen, G. Harvey, S. Orts-Escolano, R. Pandey, J. Dourgarian *et al.*, "The relightables: Volumetric performance capture of humans with realistic relighting," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 6, pp. 1–19, 2019.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[19] M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool, "Automated reconstruction of 3d scenes from sequences of images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 55, no. 4, pp. 251–267, 2000.

[20] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, vol. 7, no. 4, 2006.

[21] J. F. Blinn, "Models of light reflection for computer synthesized pictures," in *Proceedings of the 4th annual conference on Computer graphics and interactive techniques*, 1977, pp. 192–198.
[22] K. Wolff, C. Kim, H. Zimmer, C. Schroers, M. Botsch, O. Sorkine-Hornung, and A. Sorkine-Hornung, "Point cloud noise and outlier removal for image-based 3d reconstruction," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 118–127.
[23] R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural kinematic networks for unsupervised motion retargetting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8639–8648.
[24] Y. Meng, P. Y. Mok, and X. Jin, "Interactive virtual try-on clothing design systems," *Computer-Aided Design*, vol. 42, no. 4, pp. 310–321, 2010.
[25] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, "Neural volumes: Learning dynamic renderable volumes from images," *arXiv preprint arXiv:1906.07751*, 2019.
[26] M. Botsch, "Polygon mesh processing," *AK Peters*, 2010.
[27] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.
[28] S. Bouaziz, Y. Wang, and M. Pauly, "Online modeling for realtime facial animation," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 4, pp. 1–10, 2013.

Biographies

Ye-won Jang received her B.Sc. degree in Computer Engineering from Kwangwoon University in 2022, and is currently pursuing her M.Sc. degree

in Electronic Materials Engineering at Kwangwoon University. Her current research interests include 3D graphics, real-time motion capturing and 3D model animating.

Jung-Woo Kim received his B.Sc. degree in Electronic Materials Engineering from Kwangwoon University in 2024, and is currently pursuing his M.Sc. degree in the same department at Kwangwoon University. His current research interests include 3D graphics and VLSI design for deep learning.

Hak-Bum Lee received his B.Sc. degree in Electronic Materials Engineering from Kwangwoon University in 2024, and is currently pursuing his M.Sc. degree in the same department at Kwangwoon University. His research interests include multiview camera calibration for motion capture and 3D reconstruction of human motion.

Young-Ho Seo received his M.Sc. and Ph.D degrees in 2000 and 2004 from the Department of Electronic Materials Engineering of Kwangwoon University in Seoul, Korea and was a researcher at Korea Electrotechnology Research Institute (KERI) from 2003 to 2004. He was also a research professor at the Department of Electronic and Information Engineering at Yuhan College in Buchon, Korea, an assistant professor of Dept. of Information and Communication Engineering at Hansung University in Seoul, Korea, and a visiting professor at the University of Nebraska at Omaha, USA. He is now a full professor of the Department of Electronic Materials Engineering, a director of the Realistic Media Research Center at Kwangwoon University in Seoul, Korea, and a Chief Technical Officer as a Co-founder at Omotion Inc. His research interests include 3D graphics, 2D and 3D image processing, digital holography, real-time systems, deep learning for 3D data, and parallel processing.