engineering proceedings

MDPI

Proceeding Paper

Development of a User-Driven 3D Model Generation Support System and Motion Capture-Based Action Imitation †

Koharu Chiba, Homari Matsumoto, Atomu Nakamura and Shun Nishide *

Graduate School of Informatics, Kyoto Tachibana University, Kyoto 607-8175, Japan; h601026019@st.tachibana-u.ac.jp (K.C.); h601025001@st.tachibana-u.ac.jp (H.M.); h601026006@st.tachibana-u.ac.jp (A.N.)

* Correspondence: nishide-s@tachibana-u.ac.jp

† Presented at the 6th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB 2026), New Taipei, Taiwan, 24–26 April 2026.

Abstract

Recent advances in generative AI have enabled automatic 3D model generation from images. However, reflecting on the user's intentions during the editing process remains a significant challenge. Therefore, we developed a user-driven 3D model generation support system that integrates sequential image generation, 3D reconstruction, and motion capture-based animation. Building upon a previously developed incremental drawing model, user modifications are reflected throughout the generation process, and the resulting image is converted into a 3D model using T-pose normalization and Meshy AI. Furthermore, real-time motion imitation is achieved using Kinect-based skeletal tracking with scale compensation. Experimental results showed that the developed system maintained structural consistency and enabled intuitive character creation and animation, even for users without specialized expertise.

Keywords: generative AI; 3D model generation; user-driven system; motion capture; rigging; Kinect

1. Introduction

With the rapid advancement of virtual reality, augmented reality, and metaverse-related technologies, the demand for intuitive 3D content creation has increased significantly. Recent developments in generative AI have enabled automatic generation of images and even 3D models from textual or visual inputs. However, despite these advancements, a major limitation remains: users often struggle to refine generated outputs according to their intentions without relying on expert-level tools.

Traditional 3D modeling workflows require substantial technical knowledge, including mesh editing, rigging, and animation design. This creates a barrier for non-expert users who wish to create personalized avatars or characters. Therefore, there is a strong need for systems that allow users to iteratively refine generated content while maintaining consistency across multiple modalities.

In this study, we extend a previously proposed interactive drawing support system based on sequential image generation [1]. The original system enables users to modify intermediate drawing results, ensuring that the final output reflects user intentions. We expand this framework by integrating 3D model generation and motion capture-based animation to integrate image generation, 3D modeling, and animation in a unified pipeline, preserve design intent during the generation process in a user-driven framework, enable

Check for updates

Academic Editors: Teen-Hang Meen, Shu-Han Liao and Cheng-Fu Yang

Published: 21 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

real-time interaction with generated characters, and evaluate rigging consistency using established metrics.

2. Related Work

2.1. 3D Shape Generation from Images

Research on reconstructing 3D shapes from images has been widely conducted to improve the efficiency of 3D content creation. As an early mesh-based approach, Pixel to Polygon Mesh (Pixel2Mesh) [2] proposed a framework that directly generates a triangular mesh from a single RGB image by progressively deforming an ellipsoid mesh to fit the input image. This method enables a lightweight and practical 3D representation compared to voxel or point cloud approaches. However, due to the inherent limitations of single-view input, accurately reconstructing occluded regions and fine geometric details remains challenging.

To address these limitations, Pixel2Mesh++ [3] introduced a multi-view-based approach that leverages correspondences across multiple images to generate more detailed and consistent mesh models. Although effective, it requires multi-view inputs, which differs from the assumption of this study, where only a single user-edited image is available. In contrast, 3D Attribute Flow [4] improves single-view reconstruction by explicitly disentangling semantic attributes embedded in image features and incorporating them into the reconstruction process, leading to enhanced detail representation. Furthermore, recent work includes realistic image-to-3D generation via 2D and 3D diffusion synergy. Ref. [5] combines 2D diffusion models with 3D diffusion models to generate high-fidelity 3D objects and avatars from a single image. While these approaches significantly advance single-image 3D generation, they do not fully address the integration of user-edited images with subsequent rigging and animation processes.

2.2. Automatic Rigging

Rigging is an essential process for enabling animation of 3D characters, involving skeleton embedding and skinning. A classical method [6] automatically embeds a predefined skeleton into a static mesh and assigns skinning weights, allowing animation to be applied efficiently. Although this approach established a foundation for automatic rigging, it has limitations when dealing with complex geometries or large deviations from template structures. More recently, deep learning-based approaches such as RigNet [7] have been proposed, which directly predict joint positions, bone connections, and skinning weights from mesh data. Although RigNet improves generality across diverse shapes, it still suffers from inconsistencies in topology and semantic labeling. HumanRig [8] extends this line of research by introducing a large-scale dataset and a learning-based rigging framework specifically for humanoid characters, enabling more robust handling of complex generated meshes. However, its applicability is limited to human-like models. To overcome these limitations, UniRig [9] introduces an autoregressive skeleton generation method combined with bone-point attention mechanisms, enabling high-quality rigging for a wide range of character types, including non-humanoid shapes. In this study, we build upon these advances and focus on how to consistently integrate automatically generated 3D models with skeleton assignment and subsequent motion application.

2.3. Motion Capture-Based Animation

Applying motion capture data to generated 3D characters is crucial for intuitive character control. Wihardi et al. [10] proposed a low-cost motion capture system using Kinect, which transfers skeletal motion data obtained from RGB-D sensors to 3D characters in real time. This approach allows easy setup without requiring specialized equipment

https://doi.org/10.3390/engproc2026159022

or markers. However, a major challenge remains in adapting captured motion data to characters with different body proportions or scales.

Based on these previous works, we integrated three key components (3D shape generation from images, automatic rigging, and motion capture-based animation) into a unified pipeline. While existing studies typically address these components independently, our approach focuses on connecting them into a consistent workflow that reflects user-edited character designs and enables intuitive 3D avatar creation. We compared multiple automatic rigging methods, including UniRig, on 3D models generated by Meshy AI, and constructed a unified pipeline for 3D model generation and motion application that incorporates user modifications.

3. Proposed System

The proposed system integrates sequential image generation with user interaction, 3D model generation and rigging, and motion capture-based animation. The overall pipeline ensures that user modifications are consistently reflected across all stages.

3.1. Sequential Drawing Generation Model

In Ref. [1], we proposed a step-by-step drawing generation model that enables user modifications to be reflected during the drawing process. As shown in Figure 1, the model generates intermediate outputs representing the drawing process, allowing users to iteratively refine the image based on their intentions. In this study, the final generated image is used as the input for subsequent 3D model generation.

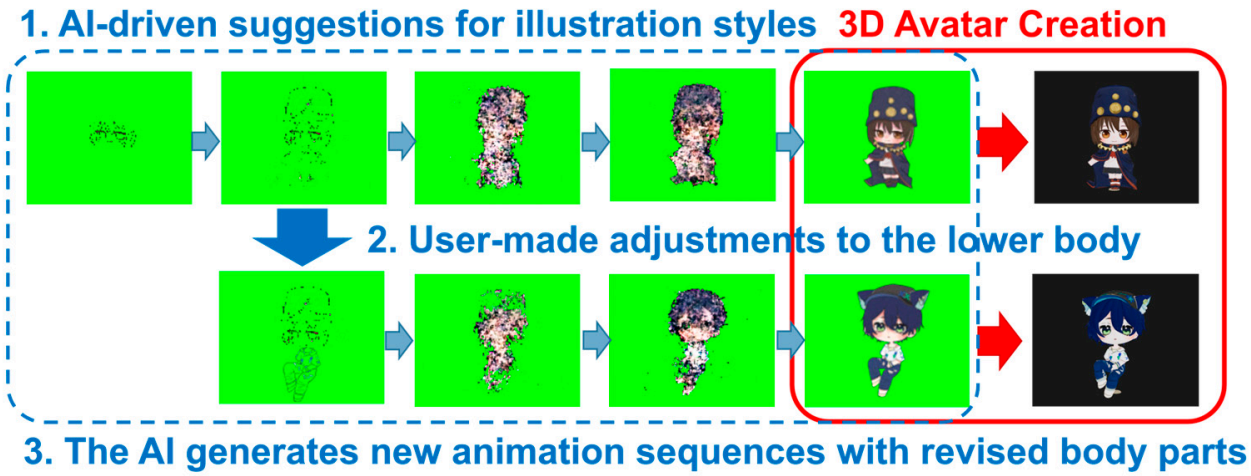


Figure 1. Overview of the proposed system pipeline. The system consists of sequential image generation with user interaction, followed by T-pose normalization, 3D model generation using Meshy AI, automatic rigging, and motion capture-based animation. User modifications are reflected throughout the pipeline to maintain design consistency.

3.2. T-Pose Normalization

The generated image reflecting user modifications (Figure 2a) is normalized into a T-pose using AI-based tools (Figure 2b) to improve the accuracy of subsequent 3D modeling and rigging processes. This transformation preserves the original design while converting the character into a standardized pose with horizontally extended arms, thereby stabilizing skeleton estimation and improving rigging performance.

https://doi.org/10.3390/engproc2026159022



(**a**) (**b**) (**c**)

Figure 2. Example of the 3D model generation process: (**a**) a user-modified character; (**b**) a T-pose normalized result; (**c**) the right image shows the generated 3D model. The figures show that the proposed pipeline preserves the original design while transforming it into a form suitable for 3D reconstruction and rigging.

3.3. 3D Model Generation and Automatic Rigging

The normalized image is then input into Meshy AI to generate a 3D model. As shown in Figure 2c, the generated results confirm that the character's features are accurately reflected in both the 3D mesh and texture. After 3D model generation, automatic rigging is performed using two methods: Mixamo (semi-automatic, markup-based approach) and UniRig (fully automatic skeleton estimation). By comparing these two methods, we analyze the rigging stability and skeletal consistency for stylized three-heads-tall characters.

3.4. Motion Capture Integration

The system uses an Xbox One Kinect sensor to capture real-time skeletal data. This data is applied to the generated 3D model within a Unity environment. A scaling algorithm is introduced to adjust joint positions according to differences in body proportions, ensuring stable animation.

4. Experimental Setup

4.1. Evaluation Metrics

In this study, we follow the static evaluation protocol proposed in the UniRig framework [9]. Three metrics are employed to evaluate the generated skeleton structures: Joint-to-Joint (J2J), Joint-to-Bone (J2B), and Bone-to-Bone (B2B). The J2J metric measures the distance error between predicted joint positions and ground truth joint positions, providing an evaluation at the joint level. The J2B metric evaluates the consistency between joints and bones, reflecting how well joint placements align with the underlying skeletal structure. The B2B metric measures the structural similarity between bones, capturing the overall topological consistency of the skeleton. For all metrics, lower values indicate closer alignment between the predicted skeleton and the ground truth skeleton. Therefore, these metrics enable comprehensive evaluation from both local (joint-level) and global (structure-level) perspectives.

4.2. Data Preparation

For evaluation, the ground truth skeleton (GT) and predicted skeleton (Pred) are defined as follows. GT is obtained from the rig automatically generated by Mixamo, which serves as a reference skeleton based on a semi-automatic rigging approach. Pred is generated using UniRig and represents the skeleton predicted by the proposed pipeline. To ensure fair comparison, all joint coordinates of both GT and Pred are normalized into a unit cube ($[-1, 1]^3$). This normalization removes the influence of scale and translation differences, allowing the evaluation to focus solely on structural discrepancies.

https://doi.org/10.3390/engproc2026159022

4.3. Experimental Conditions

To ensure the reliability of the evaluation, samples for which Mixamo failed to generate stable ground truth skeletons were excluded in advance. Specifically, samples with collapsed skeleton structures or unclear joint correspondences were removed. As a result, five samples with valid and stable ground truth skeletons were selected for evaluation. All samples share the same number of joints (22) and bones (21), ensuring consistent evaluation conditions across all experiments.

5. Result and Discussion

5.1. Results

The evaluation results based on the three metrics are shown in Table 1. Overall, the following relationship is observed: J2J > J2B > B2B. This indicates that the largest errors occur at the joint level, while the smallest errors occur at the bone structure level. In other words, although some discrepancies exist in joint positions, the overall skeletal structure is relatively well preserved.

Table 1. Quantitative evaluation results of skeleton generation using J2J, J2B, and B2B metrics for each character. Lower values indicate higher similarity between the predicted skeleton and the ground truth. # denotes the number of bones or joints.

Character ID	J2J	J2B	B2B	#Joints	#Bones
02	0.147	0.0934	0.0752	22	21
03	0.166	0.0958	0.0719	22	21
04	0.120	0.114	0.0784	22	21
05	0.193	0.189	0.131	22	21
07	0.302	0.287	0.250	22	21

Figure 3 presents qualitative examples of the generated 3D models used in the evaluation. As shown in the figure, the proposed method is able to generate diverse character models while preserving the original design features, even for stylized characters with varying shapes and proportions. These qualitative results support the quantitative findings in Table 1, indicating that despite local inaccuracies in joint positions, the generated models maintain structurally coherent skeletons that are suitable for animation and interactive applications.



Figure 3. Examples of 3D character models generated using the proposed method for ten different input characters. The results demonstrate that the proposed pipeline can generate diverse models while preserving the original design characteristics, even for stylized characters with varying shapes and proportions. The generated models are structurally consistent and suitable for subsequent rigging and animation.

https://doi.org/10.3390/engproc2026159022

5.2. Result Analysis

The relationship "J2J > J2B > B2B" is consistently observed across most samples (Table 1). This indicates that even when errors exist in joint positions, the generated skeletons are located on or near the correct skeletal structure. In other words, the developed system tends to preserve global structural consistency more effectively than precise local joint positioning.

For samples ID02–05, relatively low error values were obtained across all metrics, demonstrating high consistency between the predicted and ground truth skeletons. This suggests that UniRig can generate stable skeletons even for deformed (three-heads-tall) characters. This property is particularly important in user-driven content creation, where character proportions may vary significantly from standard human models. In contrast, sample ID07 exhibits significantly larger errors compared to other samples. This is mainly due to misalignment of some joints, likely caused by irregular mesh geometry or unique structural characteristics of the model. However, even in this case, the B2B value remains relatively low, indicating that the overall bone structure is largely preserved.

Furthermore, the observed trend that B2B errors remain relatively small even when J2J errors increase suggests that the skeleton generation process is robust in maintaining topological consistency. This behavior can be attributed to the characteristics of the UniRig framework, which prioritizes the generation of coherent bone structures over precise joint localization. These results imply that the developed system is well-suited for applications where structural plausibility is more critical than exact joint placement, such as real-time animation and interactive character control. At the same time, improving joint-level accuracy remains an important direction for future work, particularly for applications requiring high-precision motion reproduction.

5.3. Discussion

The proposed system demonstrates several advantages in terms of usability and consistency. First, it maintains high consistency between the original 2D design and the generated 3D model, enabling faithful preservation of character features throughout the rigging process. This is particularly important in user-driven content creation, where maintaining the original design intent is essential. Furthermore, the system reduces the need for expert knowledge in 3D modeling and rigging, thereby lowering the cost and complexity of character production. By integrating motion capture, the proposed pipeline also enables real-time interaction with generated characters, which enhances its practical applicability in interactive environments such as games and virtual avatars.

However, limitations remain. The performance of the system is highly sensitive to the quality of the input mesh, and joint misalignment may occur when the mesh structure is irregular or contains geometric inconsistencies. In addition, the evaluation dataset is relatively small, consisting of only five samples, which limits the statistical reliability of the results and prevents comprehensive validation across diverse character types.

To address these issues, robustness against mesh irregularities must be validated by incorporating mesh refinement or post-processing techniques. In addition, expanding the evaluation dataset and conducting comparisons with additional rigging methods will be necessary to further validate the effectiveness and generalizability of the proposed approach.

6. Conclusions

In this study, we developed a user-driven 3D model generation support system that integrates a sequential image generation model reflecting user intentions with 3D model construction from a single image. The proposed approach enables a consistent production

https://doi.org/10.3390/engproc2026159022

pipeline from 2D design to 3D model generation, allowing user modifications to be naturally reflected in the resulting 3D models. To improve rigging accuracy, T-pose normalization was introduced as a preprocessing step, aligning the character into a standardized pose while preserving its original design features. In addition, by incorporating Meshy AI, the system can generate 3D meshes and textures that maintain the visual characteristics of the input image. Furthermore, a motion capture environment using Kinect was developed, enabling real-time animation of the generated characters based on user movements. The performance of automatic rigging using UniRig was quantitatively assessed using three metrics: J2J, J2B, and B2B. The results indicate that although errors exist at the joint level, the overall skeletal structure maintains a high level of consistency. However, it was also observed that the performance is sensitive to the quality of the generated meshes, and joint misalignment may occur when geometric inconsistencies are present.

Currently, the integration of bone generation using Auto-Rig Pro is in progress, and further improvements in performance are expected through future evaluations. Future work includes comparative evaluation with additional rigging methods, improving adaptability to complex shapes, and expanding the evaluation dataset. The developed system enables an intuitive environment for users to generate and animate 3D avatars without requiring specialized knowledge in 3D modeling or rigging.

Author Contributions: Conceptualization, K.C. and S.N.; methodology, K.C.; software, K.C.; validation, K.C., H.M. and A.N.; formal analysis, K.C.; investigation, K.C.; resources, H.M. and A.N.; data curation, K.C.; writing—original draft preparation, K.C.; writing—review and editing, S.N.; visualization, K.C.; supervision, S.N.; project administration, S.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI, grant number JP23K11277.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the members of the Nishide Laboratory at Kyoto Tachibana University for their support in the development of the system and the preparation of the experimental environment. During the preparation of this manuscript and study, the authors used ChatGPT (OpenAI, GPT-5.3, accessed in 2026) for language refinement and for T-pose normalization of character images as part of the proposed 3D model generation pipeline. Meshy AI was used to generate 3D models from the normalized character images. The authors have reviewed and verified the AI-assisted outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AR	Augmented Reality
VR	Virtual Reality
3D	Three-Dimensional
J2J	Joint-to-Joint
J2B	Joint-to-Bone
B2B	Bone-to-Bone
GT	Ground Truth
Pred	Predicted Skeleton

References

1. Matsumoto, H.; Nakamura, A.; Nishide, S. Learning and Generation of Drawing Sequences Using a Deep Network for a Drawing Support System. *Appl. Sci.* **2025**, *15*, 7038. [CrossRef]
2. Wang, N.; Zhang, Y.; Li, Z.; Fu, Y.; Liu, W.; Jiang, Y.-G. Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. In *Proceedings of ECCV 2018*; Springer: Cham, Switzerland, 2018; pp. 55–71.
3. Wen, C.; Zhang, Y.; Cao, C.; Li, Z.; Xue, X.; Fu, Y. Pixel2Mesh++: 3D Mesh Generation and Refinement from Multi-View Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 2166–2180. [CrossRef] [PubMed]
4. Wen, X.; Zhou, J.; Liu, Y.-S.; Su, H.; Dong, Z.; Han, Z. 3D Shape Reconstruction from 2D Images with Disentangled Attribute Flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 3793–3803.
5. Xue, Y.; Xie, X.; Marin, R.; Pons-Moll, G. Gen-3Diffusion: Realistic Image-to-3D Generation via 2D & 3D Diffusion Synergy. *arXiv* **2024**, arXiv:2412.06698.
6. Baran, I.; Popović, J. Automatic Rigging and Animation of 3D Characters. *ACM Trans. Graph.* **2007**, *26*, 72. [CrossRef]
7. Xu, Z.; Zhou, Y.; Kalogerakis, E.; Landreth, C.; Singh, K. RigNet: Neural Rigging for Articulated Characters. *ACM Trans. Graph.* **2020**, *39*, 58.
8. Chu, Z.; Xiong, F.; Liu, M.; Zhang, J.; Shao, M.; Sun, Z.; Wang, D.; Xu, M. HumanRig: Learning Automatic Rigging for Humanoid Character in a Large Scale Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2025; pp. 304–313.
9. Zhang, J.-P.; Pu, C.-F.; Guo, M.-H.; Cao, Y.-P.; Hu, S.-M. One Model to Rig Them All: Diverse Skeleton Rigging with UniRig. *arXiv* **2025**, arXiv:2504.12451.
10. Wihardi, Y.; Rahman, E.F.; Farras, F.H.N. Kinect Based Motion Capture for 3D Character Animation. In *Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar (MSCEIS 2019)*; European Alliance for Innovation: Bratislava, Slovakia, 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.