

Digital Human Creation Based on Multi-View 3D Reconstruction: Usability Evaluation and Collaborative Workflow

Kun Fu ^{1,†}, Shuyao Shi ^{1,†}, Chengcheng Qi ^{1,†}, Xuwen Lan ¹, Pengfei Cai ^{2,*}, Zhen Liu ^{3,*}

Received 5 July 2026

Revised 25 August 2026

Accepted 19 September 2026

Published 30 September 2026



ISSN 2759-7830 (Online)
ISSN 2760-2508 (Print)

© 2026 The Author(s)
Published by Jandoo Press Co., Ltd.

This article is licensed under the terms and conditions of the Creative Commons Attribution (CC BY) license:
<http://creativecommons.org/licenses/by/4.0/>

Abstract: To create digital humans from actual people there is a practical workflow that includes capturing images, reconstructing models, fitting characters, animating faces and exchanging resources. This research has three parts; a preliminary usability testing of Metashape, creating characters from real people and developing a collaborative workflow. Six students in Digital Media Technology each had thirty minutes of instruction prior to being asked to complete seven reconstruction tasks on a 60 cm plaster bust of Voltaire. Of all possible (42) student/task combinations, thirty-four were successful which gave the study an overall task completion rate of eighty percent (80.95%). Four of those six students successfully completed the full reconstruction process resulting in a sixty-six percent (66.67%) full-process completion rate. On average, it took seventy-nine (78.75) minutes to complete the tasks. The majority of this time was spent on photographing the subject which accounted for fifty-nine percent (50.79%) of the time. Some common reasons for failures observed during the experiment included too little overlap between photographs, overly exposed areas and lost focus. An independent demonstration showed how a Metashape mesh model can be converted into a MetaHuman character that is rendered inside Unreal Engine 5. These results are used to develop a workflow that incorporates high, flat and low shooting rings for capturing images of subjects, HumanSeg as a method to remove backgrounds from images, cloud-based resource exchanges and Live Link Face as a means of providing real-time facial animation. The initial results suggest that providing guidance during image capture and checking image quality are key factors. The example of creating a character shows that creating such characters is feasible for certain use cases. Automation could improve several aspects including processing speed, visual fidelity and realism of facial movements but would need to be evaluated separately.

Keywords: Multi-view 3D reconstruction; Digital human; Usability; Photogrammetry; Metashape; MetaHuman; Collaborative workflow

Introduction

The primary problem faced by producers today, who want to use a variety of software tools to create digital humans, is how to take recognizable physical representations and turn them into editable and transferable digital entities

(Wang, & Liang, 2022; Xian, 2021). Surface reconstruction solves only some parts of this problem. Creating a recognizable head model will give you the ability to reconstruct a surface, but there is no built-in facial skeletal system, no inherent body structure, and no automatic way to link the resulting mesh with existing performance capture systems.

¹ School of Culture and Creativity, Beijing Normal–Hong Kong Baptist University, Zhuhai 519087, China; ² Yangcheng Evening News Group, Guangzhou 510000, China; ³ School of Literature and Media, Nanfang College Guangzhou, Guangzhou 510970, China.

† These authors contributed equally to this work and share first authorship

* Corresponding author. Email: caifeifei01@qq.com (Pengfei Cai); wow18782959011@gmail.com (Zhen Liu).

Application studies show why producing a usable character matter beyond its visual appearance. In museums, Mo et al. (2026) link users' adoption intentions to perceived usefulness, ease of use, aesthetic experience, and flow experience. In digital mourning, Fu et al. (2025) examine acceptance among bereaved family members of cancer patients, identifying ethical concerns and grief-related factors. These contexts give character production a wider purpose: the resulting representation enters interactions whose value depends on how people experience and interpret it.

The traditional character creation process usually follows a series of steps including sculpting, creating topography, applying textures and rigging. Each step has its own set of unique requirements for the user's artistic skill and/or knowledge of software applications. There are dedicated scanning solutions for acquiring surfaces, but the usefulness of those solutions depends on both the type of scanning equipment used, the capture environment, and the subsequent work flow required to process the images acquired.

Image-based reconstruction provides an alternative solution because capturing images using cameras or cell phones can occur quickly and inexpensively. Although accessing devices capable of taking high-quality photographs is not a problem today, making sure the images taken meet necessary quality standards prior to attempting reconstruction, ensuring adequate overlap exists between views, identifying cases where photographs have been improperly exposed or poorly focused, and integrating newly created models into other areas of your production pipeline remain significant challenges. In addressing the gap between current technology availability and production processes, the limitations were identified associated with many commercial products designed to facilitate image-based reconstruction. Therefore, our study examines the order of task execution rather than examining the individual interfaces that comprise the reconstruction process.

Our study contains three research questions:

- RQ1.** What percentage of users are able to complete the multi-view reconstruction process after receiving a short period of instruction and at what points do users experience difficulty in completing operations?
- RQ2.** In what ways can standardize data collection and automated background removal be integrated with mesh reconstruction and MetaHuman character creation?
- RQ3.** In what manner can HumanSeg processing, cloud-based collaboration via Live Link Face facial control be integrated into a single workflow and what additional factors need to be assessed?

There are three major components in our study. An evaluation of usability measures success rates and times to complete tasks for completing the sequence of reconstruction to produce a plaster bust. A live demonstration evaluates transferring a constructed mesh into an editable form. De-

sign of a workflow incorporates all elements of collecting data, automating preprocessing, creating characters, controlling facial animations, and exchanging data. All three components assess different analytical components of the study: the usability component includes six users evaluating the process of reconstructing objects; the character demonstration demonstrates one possible application scenario; and the workflow defines the structure of personnel, software and resources proposed for production.

Technical Background and Research Framework

Multi-view reconstruction as input process

Camera estimation and feature matching are the foundations of many image-based reconstruction frameworks. Incremental scale-invariant local feature extraction (Schönbberger & Frahm, 2016) and camera parameter recovery (Lowe, 2004), along with multi-view stereo (Furukawa & Ponce, 2010; Schönberger et al., 2016), for dense surface reconstruction, provide the technical context, but do not specifically detail the proprietary implementations adopted by this paper. Benchmarks (Seitz et al., 2006), and reviews on dense image matching demonstrate that a high-quality reconstruction has to be measured using defined metrics for both accuracy and completeness (Remondino et al., 2014).

Multi-view 3D reconstruction identifies spatial relationships through identifying correspondences between features extracted from images taken from multiple viewpoints. Photogrammetry studies utilizing Agisoft Photoscan represent applicable photogrammetric methodologies (Miao & Wu, 2017). Similarly, deep learning based multi-view reconstructions (Sun, B., 2022), and deep learning based multi-view facial reconstructions (Lin & Chen, 2022) discuss some of the similar technical issues.

The reconstruction environment utilized was Metashape. The Metashape work flow consists of aligning photos, generating a dense point cloud, creating a mesh, generating textures, and exporting. Although these terms identify the task assignments to participants, they do not reference the individual algorithms implemented internally, nor suggest that all stages utilize the same deep learning methodology. Neither are specific versions of software referenced, nor specific processing parameters identified in the evaluation documentation; consequently, comparisons to a specific algorithm are outside of the scope of the documented technical information. (Mohren & Schulze, 2024; Rivera & Salinas, 2025)

Quality of images is important throughout all phases because each phase depends upon the relationship between images obtained from different viewpoints. Image acquisition must produce sufficient overlap of the field-of-views so that significant features can be clearly seen in focus within the image. Simply loading photos into the software indicates that the file(s) were successfully imported into the software; it does not ensure that there exists the requisite level of

Table 1 | Components of the study and scope of assessment

Component	Event	Evidence	Scope
Usability Evaluation	Six participants; seven reconstruction tasks	Complete record and task time	Partial restoration of a plaster bust
Character Demonstration	Conversion from Metashape to MetaHuman	Production visuals and character showcase	Feasibility in a real-world case
Data Acquisition and Preprocessing Exploration	Three capture rings; background contrast	Qualitative observation of point clouds and meshes	Initial specifications and candidate interventions
Automated and Collaborative Design	HumanSeg batch processing; cloud directory	Proposed operation and asset interface	Performance and advantages not measured
Facial Drive Design	Real-time facial link connection	Designated connection and travel route	Real-time performance not measured

Note: Each component is evaluated separately.

overlap or consistence of the view to allow for successful reconstruction. Thusly, this system separates completion of operational procedures from the overall quality of the outputs produced. It is this separation that explains how completion of an initial procedure can co-exist with unsuccessful subsequent reconstruction.

From surface geometry to animated characters

Deformable facial models are used to describe the differences in appearance and shape among individuals' faces (Blanz & Vetter, 1999). Facial shapes and expressions can be created with the FLAME model using scans of an individual's face (Li et al., 2017) and a large library of facial expressions have been compiled by Cao et al. (2014) as part of the FaceWarehouse dataset.

A reconstructed surface will contain some type of geometric detail and possibly texture. A digital human that is able to animate its face needs organization beyond just the geometry of the face, which includes a way to represent a character with the ability to control the movement of its face (Cioffi et al., 2025). Character editing and asset management were performed using MetaHuman, which included all necessary pipelines for demonstrating how to edit and manage a character. Facial motion data was produced using LiveLinkFace and was demonstrated as being possible to use for facial animation.

As such, the three components of an animatable digital human - surface geometry, character assets, and motion streams - were treated independently and not as alternative ways to describe a single model. This separation also affected what could be evaluated about each component. Therefore, each transformation required its own set of evaluation criteria. For this research, only the status of completing the reconstruction task and the amount of time spent processing were quantitatively analyzed, whereas analysis comparing the quality of the final product as well as quality of the preprocessed product was qualitatively analyzed, and there was no quantitative measurement of facial driving performance.

The architecture for producing an animatable digital human divides the entire process into five separate processes: capture, preprocessing, reconstruction, character creation and facial driving. A cloud-based transfer platform allows

data to flow through these phases from one device or user to another. Capture provides photographs; preprocessing creates foreground masks and processed images; reconstruction creates point clouds, meshes and textures; character creation produces edited/rigged character assets; and facial driving creates animations based on inputted performance streams (Table 1).

Human decision-making and task boundaries

The principles to support the human-computer interaction recommend that users should be able to correct or clarify a computer function (Amershi et al., 2019). In terms of this study's evaluation, the above-referenced design principle was utilized when developing the design elements to support the inspection and recovery points in an automated segmentation process. No assessment was made as part of this study regarding the potential impact on usability of reconstruction as a result of utilizing these two intervention strategies.

In addition to viewing automation as merely delegating tasks rather than eliminating human decision-making processes, photo capture remains a manual process. Although background segmentation may be used for batch processing, it will require user review. Additionally, although reconstruction instruction can be developed in sequential format, character fitting will continue to require user verification of facial characteristics. If the platform uses independent transmission methods to manage character resource information, data transfer can also be standardized by using similar folder structure names and name conventions.

Methods

Usability evaluation setup

The usability testing took place at the School of Digital Media Art and Design at Neusoft University in Chengdu; Six students (majoring in digital media technology) participated. Participants worked with their personal mobile phones for photographing, and they utilized Metashape software that they ran from an ASUS laptop provided by the university. A 30 minute tutorial was given before testing that included basic usage of the software and workflow. Once com-



Figure 1 | Voltaire's 60-centimeter plaster bust used for reconstructing the availability assessment

pleted, participants were allowed 1.5 hours to accomplish the task.

A plaster bust of Voltaire was photographed for this project. This bust stands 60 centimeters tall as depicted in [Figure 1](#). Utilizing a rigid subject ensures no subject movement occurs while capturing images, which aids the study of how well the process works. Although using a rigid object will allow researchers to focus specifically on how well the process works, it limits the practicality of the findings for creating digital representations of real subjects because there are many factors related to living humans such as posture, facial expressions, hair and other small movements that can cause problems. Thus, testing of the bust represents a means to evaluate the reconstructed process when working under controlled conditions, rather than validate completely all aspects of digital human creation.

Seven different tasks were created: T1 – Capture images; T2 – Import images; T3 – Photo alignment; T4 – Generate a dense point cloud; T5 – Create a mesh; T6 – Texture generation; T7 – Model and point cloud export. Both the completion status of each task and how long it took were observed for recording. In this study we combined exporting into one process (and therefore only counted it once). We also excluded from our study evaluation of HumanSeg processing, transfer of the point cloud, creation of the MetaHuman characters, and use of Live Link Face with facial animation.

The participants' prior experiences using computer-aided design modeling tools, their mobile phones, laptops used in testing, number of photographs taken during the study, versions of software utilized and other factors that could have potentially influenced the time required to complete the tasks, were not documented. It was thus impossible to assess the effect of hardware on the performance of

participants under control conditions. Although all participants had completed an introductory tutorial session regarding the operation of both Blender and the Photogrammetry tools, we did not consider them complete novices. All participants were enrolled in a relevant degree program, and we will retain this variable as a potential confounding variable to affect the generalizability of the findings.

Completion measures and time analysis

A participant-task combination was either completely done (1) or was not completed (0). There were four quantitative measures of how well each participant did: individual task completion rate (the number of tasks the participant finished divided by 7); task completion rate (how many participants completed a single task divided by 6); overall task completion rate (how many of the participant-task combinations were completed divided by 42); and full sequence completion rate (how many participants completed all seven tasks divided by 6).

Each of these metrics addresses a different question. How many tasks in the task matrix were completed is indicated by overall task completion; however, how many participants completed all steps of the reconstruction process is identified by full-process completion. Participants can complete several earlier tasks and then fail to export a model. In this case, they will have increased their overall task completion count with no contribution to the successful completion of a full process.

Reporting both overall and full-process completion rates help to distinguish a participant's complete reconstruction process from their partially completed process. Time lost due to missing data is documented as "Not Applicable" (N/A) and therefore is treated as an unknown value. Time required to complete the entire sequence and time taken on average for each task are computed for participants two through six since those individuals completed all tasks. The thirty minutes allocated for training is reported separately and is not included in the computation of the above-mentioned times.

Since the purpose of this research is descriptive, the fact that sequence dependency exists among these forty-two combinations means that the forty-two combinations should be treated as non-independent observations. Failures at an early stage in a sequence could preclude the completion of later tasks. Thus, it would be inappropriate to use the data for hypothesis testing involving independent variables. Also, other relevant measures, including perception of workload, satisfaction, error in geometry, and similarity in recognizing objects, were not measured during the study and thus are represented numerically.

Live demonstration and preprocessing exploration

Another demonstration utilized an actual individual in the role of the subject for image acquisition, Metashape-based reconstructions of that acquired data using the subject's captured images, post-processing of facial features for

Table 2 | Task completion records and descriptive ratings

Participants	T1	T2	T3	T4	T5	T6	T7	Task	Rating
P1	1	1	0	0	0	0	0	2	28.57%
P2	1	1	1	1	1	1	1	7	100.00%
P3	1	1	1	1	1	1	1	7	100.00%
P4	1	1	1	1	1	1	1	7	100.00%
P5	1	1	1	1	0	0	0	4	57.14%
P6	1	1	1	1	1	1	1	7	100.00%
Total / Rate	6/6	6/6	5/6	5/6	4/6	4/6	4/6	34	80.95%

Note: P indicates participant; 1 = completed, 0 = not completed. The final completion rate was 34/42. The completion rate for the full sequence alone was 4/6 = 66.67%. A zero value downstream does not necessarily indicate independent failure.

Table 3 | Records the time and average of those who completed the full sequence task

Participants	T1	T2	T3	T4	T5	T6	T7	Total
P1	27	8	N/A	N/A	N/A	N/A	N/A	N/A
P2	40	12	3	5	8	6	2	76
P3	36	6	5	9	10	10	4	80
P4	44	15	4	6	5	7	2	83
P5	35	10	12	15	N/A	N/A	N/A	N/A
P6	40	5	6	8	8	6	3	76
Completion means	40	9.5	4.5	7	7.75	7.25	2.75	78.75

Note: The numerical units are in minutes. The method includes only P2, P3, P4, and P6 (n = 4). N/A indicates time unavailable and is not considered a zero value. Training time has been excluded.

the purposes of correction and solving in Unreal Engine 5, as well as finalizing their MetaHuman appearance edits in engine.

This test examined the possibility of integrating a recreated three-dimensional geometric representation from Metashape into the typical character creation process. As such, the subject who participated in this case study was not one of the six student subjects (as previously described) and therefore no overall completion percentage or time metrics existed for each activity. In terms of capturing images and exploring them; two additional sets of capture rings were implemented – a high ring positioned directly above the subject's head, a medium ring positioned approximately at eye-level with the subject, and a low ring positioned below eye level with the subject. A qualitative comparison was also conducted to reconstruct the same set of images taken of the subject with both background removed and not removed to assess if the presence of background information would generate undesired structure within the point clouds or cause missing sections of the face.

Results

Task completed

Participants 2, 3, 4, and 6 completed each of the seven tasks. Participant 1 completed two tasks. Participant 5 completed four tasks. This means that there were $2 + 7 + 7 + 7 + 4 + 7 = 34$ task combinations completed by participants. Overall, the task completion percentage was $34 / 42 = 80.95$ percent; additionally, the full sequence completion percent-

age for participants who had completed at least one task was $4 / 6 = 66.67$ percent ([Table 2](#)).

The completion rate for importing images was 100 % for all participants, however only five participants successfully completed the dense point cloud generation task (an 83.33 % task success), and four successfully completed the mesh generation, texture generation and exporting tasks (a 66.67 % task success). The average task completion rate per participant for the first three tasks is as follows: Participant 1 completed at a rate of 28.57 %, Participant 2 and 3 both completed at a rate of 100 %, Participant 4 and 6 both completed at a rate of 100 %, Participant 5 completed at a rate of 57.14 %. Overall, the trend from top to bottom suggests there are places in which the progression of the process came to a halt, but it does not indicate that every participant who did not complete the last two tasks attempted these tasks and was unsuccessful.

Temporal distribution

The total duration of all the recordings for those four people completing the full study was $76 + 80 + 83 + 76 = 315$ minutes or an average of 78.75 min (range 76-83). As shown in [Table 3](#), there were no instances where participants ran over the allotted 90 minutes.

The acquisition stage (T2) consumed most of the time (40 min). All other stages after acquisition (importation 9.5 min., alignment 4.5 min., creation of dense point clouds 7 min., creation of a mesh 7.75 min., texture generation 7.25 min., exporting 2.75 min.) had average times of 39.75 min. The acquisition step alone accounted for almost half of the entire time of the record ($40 / 78.75 = 0.508$); the ratio of



Figure 2 | Dense point cloud generated by participant 4



Figure 3. The reconstructed plaster bust is presented in a grid format with multiple surface views
These images provide a visual record of the reconstruction process, rather than an assessment of quantitative accuracy.

this acquisition time to the total time in the experiment is therefore called the "capture" ratio and is representative of the amount of time acquired as a whole in comparison to all available time by all subjects who successfully recorded. However, this ratio cannot provide information about how those who did not succeed distributed their available time. Also, because there are no data which differentiate between the computer processing time and manual operation time for each segment of the application, the time distribution between the two cannot be calculated either.

Participants 1 & 5 were the fastest and slowest users respectively with acquisition times of 35 and 72 minutes. Neither participant's acquisition time was replaced by a summation of the remaining segments; since both participants' acquisitions failed, they were excluded from consideration when calculating an average sequence time of 78.75 minutes, so the calculation above can be considered conditional upon completion. Thus, it is important to note that the average time of 78.75 minutes above is not indicative of a user's expected time to obtain a complete Digital Human Service.

Observed failure path

Participant 1 had a problem aligning their photos in the system. Following this, they reviewed the methods involved in acquiring the images; however, they were unable to get any two consecutive images (adjacent) to be within 50%-

60% of each other. As a consequence, the participant failed to accomplish any additional steps in the process. In addition to highlighting the connection between the capture phase of an image acquisition process and later phases involving software, this example clearly demonstrates that successful import of images into the system does not necessarily ensure that a dataset has sufficient completeness and accuracy.

The participant who produced the mesh for the mesh study described above, participant five, experienced numerous gaps throughout his mesh. His use of a number of poorly focused and overly exposed photos in addition to a large amount of blurry photos resulted in many of those images being unable to be aligned properly during the process. Additionally, he noted in the study that there were numerous areas where the structure of the point cloud was either not present at all, or was severely underdeveloped. He also noted that there were instances where structures within the point cloud were significantly out of line. These findings illustrate how poor or inconsistent input data can continue through various stages of processing to eventually manifest themselves as poor or incomplete output.

In neither of the failures identified could cause-and-effect be determined independently from one another. No measurements of overlap between images existed at the individual-image level; thus comparisons could not be made

regarding exposures of different photographs. Additionally, no complete record of events occurred during operation existed to determine whether specific levels of image quality, specific software parameters or operator decision-making processes influenced failures. Although the failure modes demonstrated here do not allow for independent assessments of causality, they clearly support moving the inspection stage up before generating grids. If the inspection occurs too late in the process, users will have already diagnosed potential problems when the evidence that supports such diagnoses would likely still exist within the images captured.

Preprocessing and visualization output

The outputs produced by participants two and four included more detailed point clouds with a higher level of detail as well as preserved object characteristics and texture. Participant 2 utilized light room to adjust both color and exposure during the process of importing images into the software. Similarly, participant 4 used the automated background removal tool within the same software package. In comparison to the results of participants 3 and six their results were superior in regards to object details and quality of output ([Figures 2 and 3](#)).

These were all qualitative observations. The visual evidence did not allow for a standardized count of points, a balance in capturing objects under different lighting conditions, nor an ability to correct geometric errors. Thus, while a denser point cloud does not mean that it will be a better representation of the object being reconstructed, no one can isolate how different operations (preprocessing) contributed to the differences seen in the output. The images provided are of the results and reasoning behind the preprocessing analysis (as opposed to quantifiable improvements due to Lightroom or segmentation operations). In comparisons using human subjects on their own, when reconstructing models without separating the foreground from the background, many people confused areas of facial point clouds with other areas of a similarly colored background region, which resulted in local holes within the facial mesh. Therefore, removing the background may be an additional step that would improve upon some workflows. However, there is still much work needed to quantify this improvement, validate whether it is repeatable, and compare HumanSeg against manual masking; issues that have yet to be addressed.

Integrated Digital Human Creation Workflow

Use three capture rings for standardized collection

An alternative to using a fixed spacing between the rings would be to use variable spacings such that some of the rings will have less than 20 degrees of spacing while other rings will have greater than 40 degrees. In this way, we could

test if the overlap is sufficient across all possible orientations of the face. In addition, even though the three ring arrangement provides substantial redundancy to help solve many of the problems inherent in reconstructing a 3D image from 2D photographs, there are no guarantees that there will be sufficient overlap in all cases. Similarly, if your subject wore glasses or a hat with brims, then they might occlude a portion of their eyes, nose and mouth from the camera. Since these occlusions would be common across all views from both sides of the subject, then additional cameras positioned at different angles would provide redundant information about occluded regions.

However, since it is impossible to predict where an occlusion might occur on any given day for any given person, then you cannot rely solely on the assumption that a particular set of camera angles will allow you to see everything that needs to be seen. That is why I emphasized earlier that a verification step is needed. Additionally, although the three-ring plan allows us to verify that the requirements for 50% to 60% overlap between adjacent images are met, it does not replace them.

HumanSeg background preprocessing

Preprocessing involves using the HumanSeg model from PaddleSeg's open-source environment to create batch segmented foreground. It then uses a python-based pipeline to read through the image capture directory, generate foreground masks, and save resulting processed images. All three images; original, masked, and processed, should be retained together so operators have access to all the evidence required to assess whether segmentation resulted in loss of relevant data upon reviewing abnormal reconstruction results.

There exist multiple types of segmentation models within PaddleSegs HumanSeg designation, as opposed to one specific version (DeepLabv3+) ([PaddlePaddle, 2024](#)). In order to ensure a reproducible version of this process there must be specified what model was used, what pre-trained weights will be used, the input resolution size, and which inference device will be utilized. The primary purpose of this process is to remove the background from images while selecting which model to use as an explicit step.

Prior research suggests batch segmentation as an approach; however, none of prior research provided executable code, model configuration details, processing log files, or performance evaluation metrics that would allow for assessment of anticipated improvement in terms of processing speed, accuracy, or reconstruction time.

Areas requiring reconstruction include the ear, jawline, hairline and scalp region. Removing incorrect amounts of foreground could produce either missing segments or altered appearance contour segments. Additionally, if too much background is retained it may render matching problems unresolvable. As such, the proposed method includes an additional manual review component where mask corrections/re-captures may be made when necessary. Manual reviews were optional vs. mandatory during testing with six

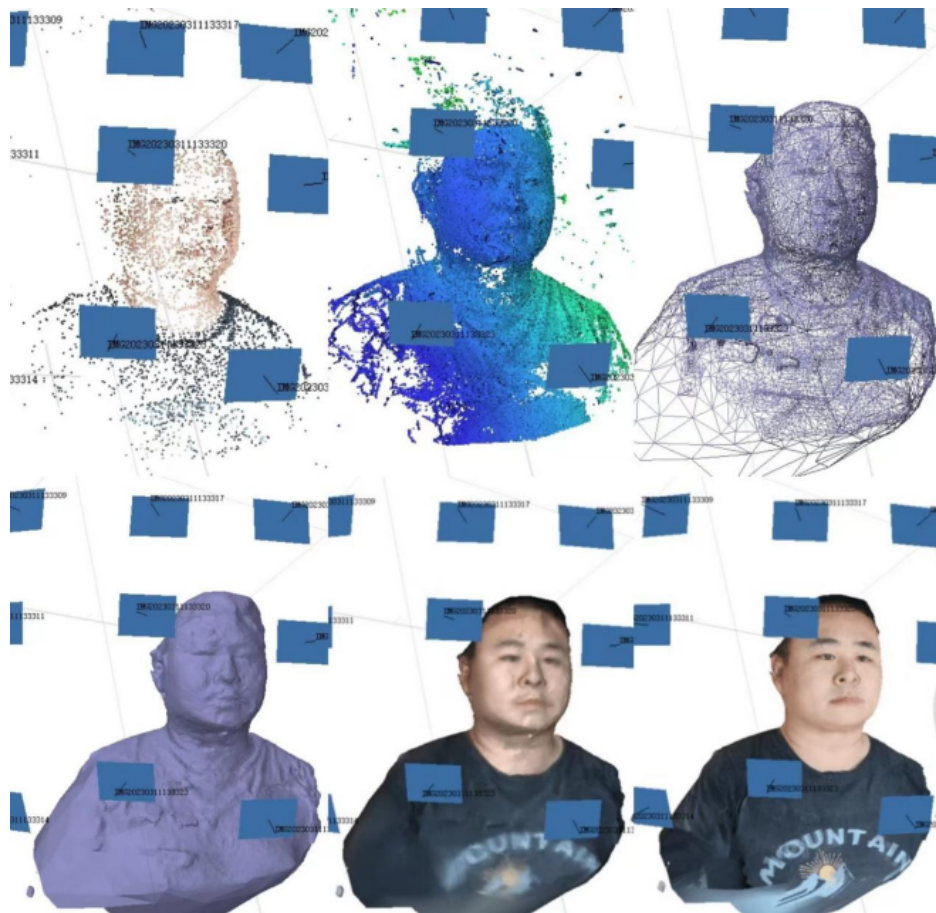


Figure 4 | Intermediate Metashape output results for live demonstration, including point cloud, mesh, and textured representations

participants. Both background removal and exposure adjustments should be evaluated independently since they represent two distinct elements of the input and potentially interact with shooting conditions.

Reconstruction and asset preparation in metashape

Reconstruction involves importing images, aligning those images, creating a dense point cloud, eliminating unnecessary points if they exist, mesh construction, adding texture, and exporting the desired files. It should be noted that cleaning a point cloud represents one step in a pipeline process, as opposed to being an eight distinct task for users in terms of usability metrics. In doing so, it provides the same production specification, and thus allows for consistent evaluation and clarity with regard to potential quality assurance issues.

Each checkpoint follows the software generated representations of each stage. Following image alignment, the operator reviews both the input images and associated camera data for validity. Upon completing the dense reconstruction, the operator examines areas where there may have been missing or misaligned features. When reviewing the mesh the focus will be upon whether the face area has continuity

for use by character artists. None of these steps represent establishing an unmeasured threshold of precision; instead they are used to identify which evidence must be collected prior to passing on the output.

The live demo shows many examples of reconstructed objects from the perspective of multiple views - e.g., point clouds, meshes, and textured models (Figure 4), all of which show the progress of software through time.

Export preparation must also maintain the relationship between the mesh and the texture data it has been paired with. Point clouds, surface meshes and image textures are all generated from the same photographs, but they have different functions. The input to the MetaHuman pipeline will be a mesh that can be imported into the application along with any other required data, as opposed to being an unassociated point cloud model.

MetaHuman fitting and appearance editing

The reconstructed mesh was imported into Unreal Engine 5 as part of the demonstration. This involved importing a selected frame from the animation into the engine to adjust facial features. Facial features were adjusted using the engine's MetaHuman pipeline to identify a character. Appearance edits were made in the cloud-based interface used



Figure 5 | MetaHuman facial fitting and appearance editing within the shown toolchain
The panel displays character representations based on intermediate results and presets.

with the associated workflows. Assets for the characters were also obtained through Quixel Bridge and rendered within the game engine. Sun (D.) (2022) references are provided to detail additional information about the pipeline utilized during the production. Each stage of the demonstrated sequence is followed in the order they occurred.

As illustrated in [Figure 5](#), this demonstrates how to perform both facial fitment and customize your character's appearance. The picture represents a centralized facial representation that can be edited according to pre-set appearance options. Allowing users to correlate their scans' geometric data with an editable, rigged character for further usage. The final asset is generated by both reconstructed input data and fitting process, presets available to you and user-defined choices; therefore, the final product will never be a complete replica of your original scan or completely automated replication of textures.

The character was also created as an Unreal Engine 5 (UE5) asset. As seen in [Figure 6](#), the example shows how the final asset is visually represented. However, it does not detail the specific techniques that were utilized to create each pose shown, nor does it include measures for the accuracy of either the facial capture or live link face performance data. Fa-

cial expressions visible in this representation could have been achieved through many different types of animations; therefore, their visibility alone cannot be used to confirm whether a real time facial capture process has occurred.

This was an example of how you do not need to utilize specialized 3D scanning technology, green screens, or different modeling systems for Maya or 3ds Max. In this example it outlines the tools used within the context of the project and does not imply that each subject will require some form of baseline (minimum) equipment. Additionally, as there are no fully detailed production log entries available, nor similar production environments to compare to, we cannot accurately quantify any time savings from fast character edits.

While the tools outlined above reflect how assets were created, plug-in interfaces, cloud-based editing capabilities, and asset delivery methods must be compared with those utilized when reproducing. This outline is not intended to verify the equivalency of any current interface(s), nor provide regionalized MetaHuman services. The scope of supportable conclusions is limited: Only a reconstructed human mesh has been linked to both character creation and engine presentation; Perceptual likeness, Geometric fidelity, and Repeatability Success have not been quantified.



Figure 6 | Character poses and facial expressions are rendered using Unreal Engine 5

These examples do not specify a particular motion capture source, nor do they establish a Live Link Face performance.

Real-time facial driving

Following the creation and binding of the respective character, Live Link Face becomes the facial motion input module to be integrated into the system. To create the necessary connection, the smart phone and computer share the same local network. The computer's IP address is registered as a target at the smart phone, and in the project settings in Unreal Engine, the incoming video capture source is selected. This stream is then linked to the respective character, such that movement of the face during image capturing (photo) can influence animation of the character.

While full body capture and autonomous interaction result in mapping of a user's entire body to a digital representation and subsequent interactive behavior of the digital representation based solely on user actions, facial driven functionality results in mapping of a user's facial expressions/behavior to a digital representation. Because latency, frame rate, expression sequence and tracking error data were not reported by the study, expressing real time operation only represents a desired operational objective rather than an actual measureable low-latency performance. As a result, these features were excluded from the completion rates of the 6 subjects who participated in completing their respective reconstructions.

Cloud collaboration and responsibility allocation

A key objective of the workflow is to reduce unnecessary handoffs and ensure that each party knows what to do with their input/output, and when to expect verification of completion. Manual capture is retained to prevent the need to

require electric camera tracks or similar specialized equipment to support the light weight workflow. Our usability assessment did not examine alternative ways to divide up the workflow.

We designed our Cloud Directory to organize items by topic and included all relevant photos taken during the session, pre-processed versions of those images, reconstruction output files, and exportable character project data. We maintained both capture ring numbers and sequence identifier labels. Task status log entries were added to provide receiving operators with a way to confirm that necessary files are complete prior to using them. Two separate storage services exist at the service level: image capturing services and model/generating character services. All three steps of processing (pre-processing, reconstruction, and character generation) occur between these two.

Cloud Exchange alone provides no clear indication of whether there are any time advantages. Several factors including but not limited to; file upload times, integrity of assets once uploaded, permission levels of accessing the assets, and correction of intermediate output artifacts will determine how effective collaboration will be. Therefore [Table 4](#) illustrates the objects involved in the handoff process and required evaluation, not a validated production service.

Discussion

Data collection quality as an availability issue

The four students who finished the Avatar reconstruction within the time frame allowed (ninety minutes) showed

Table 4 | Workflow handover and follow-up evaluation requirements

Stage	Input and Output	Operation	Assessing Priorities
Acquisition	Group avatar photos	Manual collection with recommended guidelines	Coverage, Focus, and Re-capture
Preprocessing	From photos to masks and processed images	Batch processing of human segmentation with manual review	Foreground loss and correction time
Reconstruction	Generating point clouds, meshes, and textures from images and masks	Metashape Sequence and Inspection	Alignment, completeness, and timing
Character Creation	Importable mesh to rigged character asset	Function confirmation, resolution, and editing	Similarity and Adjustment Effort
Facial Driving	Convert characters and capture streams into animated demonstrations	Real-time facial link connection and configuration	Connection successful, latency, and tracking error
Cloud Collaboration	Traceable handover of stage assets	Topic Directory and Proposed Version History	Waiting for transmission and handover errors

Note: This is a workflow proposal, not an end-to-end system evaluated in a six-person research study.

that students with some prior knowledge of computer graphics could use this local workflow. The reason that two students failed was because of how they prepared their data for the software. One student did not get enough overlap on his images; another student was unable to capture quality images. This indicates that usability is greater than simply using menus correctly; it involves creating correct input to produce acceptable output. Therefore, although a user might be able to navigate the interface and import incorrect images, a guide to acquire consistent images could be as valuable as improving the navigation of the menu. While this study did not determine whether or not such guidelines would improve performance, it determined areas where interventions can take place and what should be measured when determining those outcomes.

Acquiring images accounted for more than half of the total average number of steps taken for each user to successfully complete the task. Hence, if we can assist users in obtaining images, we may be able to obtain a reasonable amount of improvement in terms of optimizing their ability to accomplish tasks. Additionally, there are many other ways to optimize their work flow including increasing the speed at which they process their acquired images.

Achieving performance metrics and successful service delivery

In terms of usefulness for customers who want usable roles, there is a distinction between the usability of a workflow which only performs the first few steps versus a workflow which performs all the necessary steps. While both metrics are useful, the importance lies in defining what they measure by indicating the denominator. The former measures how much progress was made through the intermediate stages, whereas the latter assesses the percentage of cases where the process resulted in a model ready for use at the reconstruction endpoint.

Reconstructing an endpoint represents less than half of the tasks included within the proposed service. Generating a character with verified facial-driven capability is a much

broader concept than simply exporting a bust. It also includes additional phases including fitting, adjusting, transferring and animating that can include many additional stages to fail during these steps. Therefore, when evaluating services, it is important to define success based on deliverables and evaluate the outcome after each stage of transformation.

Automation and the redistribution of effort

HumanSeg, batch processing, ordered reconstruction commands, and cloud directories are all addressing other forms of repetitive processes. They reduce the amount of time spent on each image, optimize how the software performs its process, and standardize where data can be stored and retrieved. Unless you factor in the remaining manual labor needed after using each method, you cannot just add the advantages together.

If the segmentations are imperfect, the user will likely spend an equal if not more amount of time inspecting and correcting them instead of completely eliminating the need for human input. In addition, while some parts of the workflow may retain the ability to allow human decision-making based on what is required as a representation for future steps in the workflow, there are many steps that do not. A surface may technically be able to be imported, however, it would likely need a great deal of additional modification to reach the desired level of similarity. Cloud-based collaboration also transfers work. By separating acquisition from reconstruction, users can take advantage of their own hardware and expertise. However, this creates new transfer issues. The recipient operator needs to know who the subject in question is, whether acquisition is complete, and which version of preprocessing was last used. Although traceable assets aid in coordinating activities, they have not been proven to either decrease error rates or decrease wait times. Automating the benefit provided by automation should be limited to specific operational areas and measurable results, rather than being dependent upon only having cloud storage.

Portability and visual fidelity

Moving from a plaster bust to a real human being is a significant methodology transition. While the bust is stationary, the live subject has the potential to be moving or have varying expressions as you capture photos. In addition, variations in the subject's hair and other visual surface features can impact your results when using a plaster bust. Fixed object assessments are unable to take into consideration any of these issues. Demonstrating tools is simply one of many facets that demonstrate how users interact with tools and will never replace repeated usability testing done by actual living humans.

The evaluation of visual realism will be assessed from three different viewpoints: geometric correspondence to the subject, perceptual similarity between the subject and the character, and aligning expressive qualities of facial animations. A character developed from a pre-defined model may look realistic, however it may differ from the subject; similarly, high static similarity does not mean that expressions will be accurately transferred. Images like those shown above assist in illustrating the various stages of production and do not provide quantitative levels of fidelity.

There was no comparison with either modeling techniques or specifically designed scanning equipment. This is because in this instance we chose to use standard photography instead of a scanner as our alternate process; however, this is not always cost effective. Software licensing costs, computing resources, cloud operations, etc., corrections were not calculated. As such, cost-effectiveness and ability to reduce technical expertise requirements continue to be major design considerations. However, cost comparisons and learning demands need to be directly investigated as well.

Priorities for subsequent verification

Acquisition validation includes validating proposed parameters using documented conditions. All acquisition-related items that are to be tested must be identified including but not limited to; number of subjects, camera model, light source(s), number of shots taken, subject composition, image overlap, and alignment participation, grid orientation and integrity, and the need to take retakes. Identifying all of the above factors will assist in assessing the three-ring process as a method of obtaining facial data instead of adopting it based on its easy to describe configuration.

Pre-processing validation involves applying different processing conditions to the same set of images. Examples of pre-processing conditions would include comparing the original, un-processed images with manually created mask images and specifically HumanSeg processed images. These comparisons may include measures such as how well each condition preserved the foreground region of interest, how long each condition took to process, and what was the quality of the reconstructed image.

The End-to-end studies included evaluating data collection, data upload, segmentation, reconstruction, character fitting, and facial driving within one single production environment. Time stamps should be collected individually for

manual processes, automated calculations, data transmission delays, and for recovering from failed connections. Both experienced users and inexperienced users should participate in testing. Prior to testing success criteria must be established. Satisfaction and workload can no longer be measured by only measuring completion rates; therefore new methodologies must be developed to measure user satisfaction and workload. Evaluation of Facial Driving Performance includes assessing facial expressions through recorded sequence of expression, multiple attempt to connect, measured times of successful connection, and assessing motion correspondence.

Limitations and implementation requirements

The evaluation had three primary disadvantages: there was only 6 participants and the time until failures occurred were incomplete, and the pre-processing methods used were not consistent. Details about hardware components, configuration settings, and the number of scans conducted were not provided and thus reduce the ability to replicate the study as well as to analyze how variations in these variables impact the results. Demonstrations of characters, as well as background contrast, lacked repeat trials and quality metrics for assessing the quantity of each; they should be employed for evaluating processes, determining areas of interest for further experimentation and developing hypotheses. Functional services require that it is possible to clearly document all required version(s) of utilized software, required interface(s), licensing terms, and permission(s) to handle image data. Storage on cloud systems must have access controls and deletion protocols to ensure that only authorized users utilize photos, grid(s), and role item(s). These represent deployment specifications that are distinct from those that demonstrate mechanism verification within the scope of this assessment. They make the specification complete while preserving an identification of mechanisms for development and those that demonstrate performance.

Related work on brain-computer interfaces in elderly care connects human-machine trust to privacy, informed consent, autonomy, and communication of uncertainty (Fu et al., 2026). For the person-specific workflow proposed here, this perspective supports explaining how captured images and derived character assets will be used, who can access them, and which changes arise during fitting and animation.

Conclusion

This research combined usability analysis of multi-view reconstruction with live role demonstrations and shared production processes. A total of 6 participants participated in a 7-task assessment. The overall task completion percentage was 80.95%; however, the total sequence completion percentage was only 66.67%. Of the 6 participants who successfully completed all 7 tasks, their average completion time was 78.75 minutes; in particular, the time spent collecting data (i.e., 50.79%) represented the largest portion of the

participant's total completion time. Analysis of both unsuccessful sequences as well as the time each participant took to complete each step also indicates that providing instruction on how to collect data, and conducting early input checks will require additional development.

The MetaHuman demo illustrates a clear example of how human mesh reconstructions generated by Metashape are able to integrate into editable character generation, and engine based rendering. Furthermore, this process expands upon the prior established pipeline using a consistent capture configuration, HumanSeg pre-processing, cloud-based resource sharing, and Live Link Face facial control technology. However, these elements have yet to create a completely automated process. Therefore, additional testing is required to evaluate its end-to-end usability, automate some aspects of its functionality, establish a similarity level between the subject being analyzed, and determine if it adequately simulates the individual(s) intended to be animated.

Funding: Zhuhai Social Sciences Planning Research Project (No. 2025GJ057)

References

- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3290605.3300233>
- Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99*, 187–194. <https://doi.org/10.1145/311535.311556>
- Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2014). FaceWarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3), 413–425. <https://doi.org/10.1109/tvcg.2013.249>
- Cioffi, F., Masullo, M., Pascale, A., & Maffei, L. (2025). Speech intelligibility in virtual avatars: Comparison between audio and audio-visual-driven facial animation. *Acoustics*, 7(2), 30. <https://doi.org/10.3390/acoustics7020030>
- Fu, K., Li, P., Peng, X., Li, S., Zhao, Y., Chen, X., Ding, Y., & Ye, C. (2026). From technological closed-loop to human-machine trust: a systematic review of ethical and communication challenges of brain-computer interface in elderly care. *Frontiers in Digital Health*, 8, 1877688. <https://doi.org/10.3389/fdgth.2026.1877688>
- Fu, K., Ye, C., Wang, Z., Wu, M., Liu, Z., & Yuan, Y. (2025). Ethical dilemmas and the reconstruction of subjectivity in digital mourning in the age of AI: an empirical study on the acceptance intentions of bereaved family members of cancer patients. *Frontiers in Digital Health*, 7, 1618169. <https://doi.org/10.3389/fdgth.2025.1618169>
- Furukawa, Y., & Ponce, J. (2010). Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8), 1362–1376. <https://doi.org/10.1109/tpami.2009.161>
- Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6), Article 194. <https://doi.org/10.1145/3130800.3130813>
- Lin, L., and Chen, J. (2022). Research on multi-view 3D face reconstruction methods. *Video Engineering*, (7), 203–207. <https://doi.org/10.16280/j.videoe.2022.07.050>. [In Chinese; title translated.]
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110. <https://doi.org/10.1023/b:visi.0000029664.99615.94>
- Miao, D., and Wu, J. (2017). Image-based 3D reconstruction and accuracy assessment using Agisoft Photoscan. *Engineering of Surveying and Mapping*, (8), 41–44. <https://doi.org/10.19349/j.cnki.issn1006-7949.2017.08.009>. [In Chinese; title translated.]
- Mo, J., Chen, H., Ye, C., Wang, Z., & Chen, C. (2026). Exploring the drivers of users' adoption of museum digital humans. *npj Heritage Science*, 14, 43. <https://doi.org/10.1038/s40494-026-02313-0>
- Mohren, J., & Schulze, M. (2024). Automated cleaning of tie point clouds following USGS guidelines in Agisoft Metashape professional (ver. 2.1.0). *MethodsX*, 12, 102679. <https://doi.org/10.1016/j.mex.2024.102679>
- PaddlePaddle. (2024). PaddleSeg PP-HumanSeg documentation, release/2.6. <https://github.com/PaddlePaddle/PaddleSeg/tree/release/2.6/contrib/PP-HumanSeg>. Accessed October 13, 2024.
- Remondino, F., Spera, M. G., Nocerino, E., Menna, F., & Nex, F. (2014). State of the art in high density image matching. *The Photogrammetric Record*, 29(146), 144–166. <https://doi.org/10.1111/phor.12063>
- Rivera, J., & Salinas, P. (2025). Workflow for photogrammetric digitization using Agisoft Metashape. *Figshare*. <https://doi.org/10.1371/journal.pone.0336996.s003>
- Schönberger, J. L., & Frahm, J. M. (2016). Structure-from-motion revisited. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4104–4113. <https://doi.org/10.1109/cvpr.2016.445>
- Schönberger, J. L., Zheng, E., Frahm, J. M., & Pollefeys, M. (2016). Pixelwise view selection for unstructured multi-view stereo. *Computer Vision – ECCV 2016, Lecture Notes in Computer Science*, 9907, 501–518. https://doi.org/10.1007/978-3-319-46487-9_31
- Seitz, S., Curless, B., Diebel, J., Scharstein, D., & Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06), 1, 519–528. <https://doi.org/10.1109/cvpr.2006.19>
- Sun, B. (2022). Multi-view 3D reconstruction based on deep learning. Master's thesis, Dalian University of Technology, Dalian, China. [In Chinese; title translated.]
- Sun, D. (2022). An introduction to MetaHuman principles and workflows. *Zhihu*. <https://zhuanlan.zhihu.com/p/445646111>. Accessed May 5, 2023. [In Chinese; title translated.]
- Wang, Y., and Liang, X. (2022). Virtual subjectivity and virtual materiality: On digital human representations. *Journal of Fujian Normal University (Philosophy and Social Sciences Edition)*, (5), 49–58, 170. [In Chinese; title translated.]
- Xian, H. (2021). A study of Byung-Chul Han's aesthetic thought on digital humans. Master's thesis, Lanzhou University, Lanzhou, China. [In Chinese; title translated.]