electronics

MDPI

Article

Non-Rigid Point Cloud Matching Based on Invariant Structure for Face Deformation

Ying Li [1], Dongdong Weng [1,2,*] and Junyu Chen [3]

[1] Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China
[2] AICFVE of Beijing Film Academy, 4 Xitucheng Rd, Haidian, Beijing 100088, China
[3] Advanced Research Center for Digitalization of the Traditional Drama, The Central Academy of Drama, 39 Dong Mianhua Hutong, Dongcheng, Beijing 100710, China
* Correspondence: crgj@bit.edu.cn

Abstract: In this paper, we present a non-rigid point cloud matching method based on an invariant structure for face deformation. Our work is guided by the realistic needs of 3D face reconstruction and re-topology, which critically need support for calculating the correspondence between deformable models. Our paper makes three main contributions: First, we propose an approach to normalize the global structure features of expressive faces using texture space properties, which decreases the variation magnitude of facial landmarks. Second, we make a modification to the traditional shape context descriptor to solve the problem of regional cross-mismatch. Third, we collect a dataset with various expressions. Ablation studies and comparative experiments were conducted to investigate the performance of the above work. In face deformable cases, our method achieved 99.89% accuracy on our homemade face dataset, showing superior performance over some other popular algorithms. In this way, it can help modelers to build digital humans more easily based on the estimated correspondence of facial landmarks, saving a lot of manpower and time.

Keywords: digital humans; non-rigid deformation; correspondences of landmarks; shape context; consistent topology

Citation: Li, Y.; Weng, D.; Chen, J. Non-Rigid Point Cloud Matching Based on Invariant Structure for Face Deformation. *Electronics* **2023**, *12*, 828. https://doi.org/10.3390/electronics12040828

Academic Editor: Stefanos Kollias

Received: 30 December 2022
Revised: 28 January 2023
Accepted: 1 February 2023
Published: 6 February 2023

1. Introduction

Creating high-precision and ultra-realistic digital humans is not only widely sought after in the gaming and film industries, but also increasingly popular in AR/VR applications [1,2]. Although automatic single-view avatar digitization solutions are available [3–9], in pursuit of high fidelity, professionals still choose multi-view stereo (MVS) techniques for 3D face reconstruction [10–12], which needs to capture multi-view photos of real faces with markers [13–15]. Then, the camera calibration module decodes the 2D information of the captured facial feature points to build a 3D face model [16–18].

However, the face model reconstructed using this method always contains a large number of vertices, which is computationally intensive, making it difficult to perform subsequent steps such as bone binding and animation driving [19]. Therefore, an appropriate topology needs to be rebuilt for it. Moreover, in order to make the digital avatar more realistic, it is necessary to scan various expressions of the target actor as a reference, and these expressive models need to maintain a consistent topology [20]. A connection must be established, which can be used to warp an ideal model to each scan to ensure topology consistency [19]. In this process, optimization usually relies on art modelers manually labeling hundreds of corresponding feature points for each scan. Due to the heavy workload, in the current asset generation, the completed rig of a character that includes different expressions takes a long time to finalize [21,22]. It is especially important to automatically find the correspondence between deformable models in demand, which is also the focus of the paper.

Facial landmarks skip in pace with the change in expression. They can only be obtained with the non-rigid registration method instead of the tracking algorithm because of non-sequence [23]. In existing research, the iterative closest point (ICP) [24] searches for the nearest target point to the source point set one by one as pairs, but it fails in point sets with large deformation. Thin plate spline robust point matching (TPS-RPM) [25] evaluates correspondence through distance, softassign [26] and deterministic annealing [27]. Coherent point drift (CPD) [28] combines distance and motion coherent theory (MCT) [29] to solve the matching probability matrix with the maximum likelihood method. Although MCT improves the stability of spatial transformation in registration, it does not play a role in the sparse points on the face but rather limits the magnitude of landmark movement to facial expressions.

ICP, TPS-RPM and CPD only employ the Euclidean distance as the global feature to evaluate correspondence [24,25,28] without describing local features in a point set, leading to insufficient accuracy when global features are similar.

Robust point matching-preserving local neighborhood structures (RPM-LNS) [30], robust point matching algorithm based on L2E (RPM-L2E) [31] and preserving global and local structures (PR-GLS) [32] use a single shape context (SC) feature [33] as descriptor of the point set, which makes them unable to distinguish local structures when they are similar, leading to regional cross-mismatch during registration.

To this end, we propose a non-rigid marker matching method to break the limitation mentioned above and solve correspondence automatically between the neutral model and other expressive face models. In this way, it could be possible to implement batch reconstruction and batch re-topology based on such correspondence, so that subsequent processes such as bone binding and driving are efficient. Our contributions are listed as follows:

- We explore the properties of the texture space and normalize the global structure features of expressive faces based on added invariant topology, which decreases the magnitude of variations in face movements.
- We make a modification to the traditional SC feature to solve the problem of regional cross-mismatch and combine the modified SC and European distance features to improve the effectiveness of feature description.
- We built a photo dataset of real persons with face deformation. Experiments were carried out on real data with extreme expressions and a high percentage of outliers, demonstrating the capabilities of our method.

The remaining parts of the paper are organized as follows: Section 2 describes the overview and technical details of our proposed approach. Section 3 presents the experimental protocol and assesses the results. Section 4 concludes this study, highlighting its contributions and limitations and proposing future work.

2. Method

In this paper, we propose an automatic facial marker matching method based on non-rigid point set registration. Given a set of expressive faces with artificial markers obtained by scanning real persons, point information is extracted using a marker detector for different face images, while the grid attached to the surfaces is detected by the mesh predictor. Then, grid and point information is mapped to the texture space at the same time. Because of the consistent topology of the grid, the points can be roughly positioned after texturing. We call this step coarse positioning. Next, an additional cross corrector is incorporated into the SC feature descriptor for further fine matching, providing point set correspondences for the scanning model to reconstruct the topology. Figure 1 shows the overview of the proposed method.

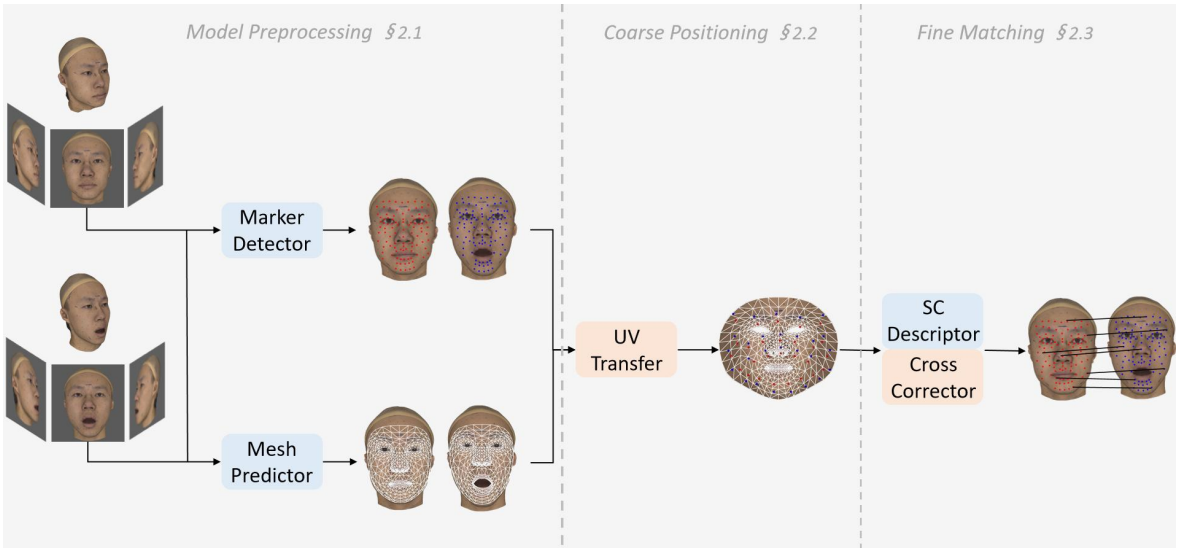

Figure 1. Overview of the proposed method.

2.1. Model Preprocessing

Our task is to find the correspondence of the markers on different expression models, which are obtained from 3D scans of real faces with markers. In order to simplify the problem, we firstly render the model with different viewpoints to obtain 2D images. After matching the markers in expressive images for each single view, we project the markers onto the 3D model using the collision detection algorithm [34] and finally realize the matching solution of the 3D sparse point cloud.

We first adjust the scan model to the same scale based on rigid ICP [24] and then use the Blob detector of OpenCV to extract 2D marker information from the images, which are rendered with the same viewpoint using 3D rendering software. The face meshes are predicted by MediaPipe Attention Mesh [35]. Then, the unnecessary points are automatically cleaned up by means of area limitation of the mesh, as shown in Figure 2. Finally, the meshes and the point sets with redundant points manually removed are used as the inputs of the subsequent matching algorithm.

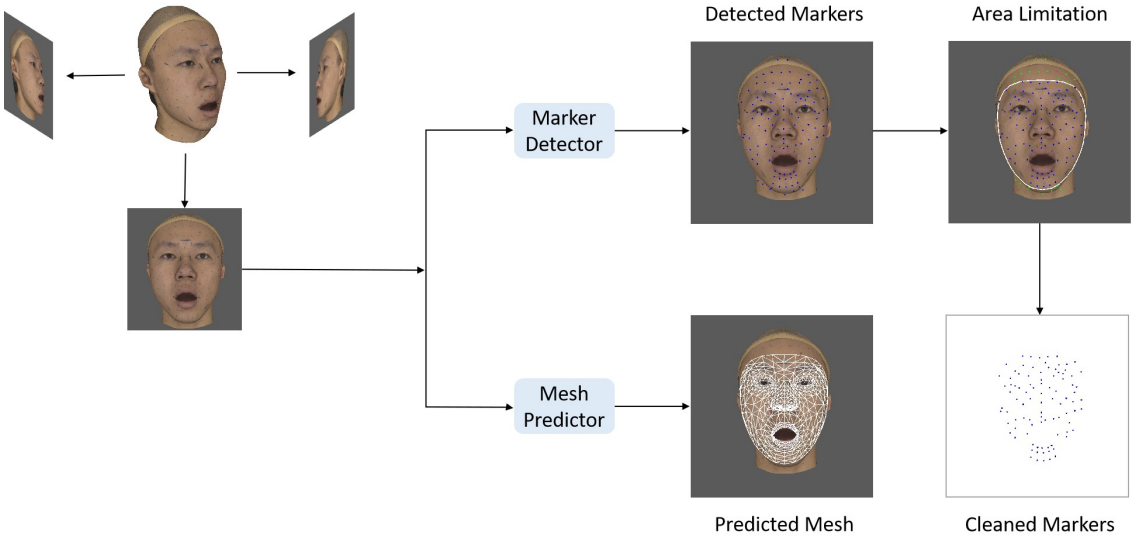

Figure 2. Overview of preprocessing.

2.2. Coarse Positioning

In this part, we propose a shape context descriptor based on a UV coordinate system (SC-UV), which realizes coarse positioning by mapping from the image space to the texture space.

Texture is the basic unit to save the surface information of 3D objects and generally works in the form of 2D images. In the process of model rendering, the mapping from the 2D texture space to the 3D object space can be realized based on the recorded information of details on the object surface.

In this paper, even though the mesh provided by the MediaPipe tool is pseudo-3D [35–37], we can still use this mapping idea to establish the mapping principle between the original image space and the 2D texture space. Based on this principle, we can realize the conversion between the original XY coordinate system and the UV coordinate system.

Because Attention Mesh [35] has a unified topology on different faces, it has the same position in UV coordinates in the texture space. In this way, we can calculate the position of each pixel near the grid based on the relative relationship. This is the basis for the implementation of our proposed approach.

After obtaining the marker and mesh information of expressive faces in preprocessing, we transfer it to the texture space at the same time. Importantly, the markers on expressive faces share correspondence in a consistent topology of the mesh, which is fixed in the UV coordinate system. The specific mapping principle is reported below.

As shown in Figure 3, given a target point $P = (x, y)$ of the image space, which is located in triangle $Mesh = \{A, B, C\}$, the position of P can be represented by a weighted sum of the positions of the vertices.

$$P = \alpha A + \beta B + \gamma C \tag{1}$$

where A, B and C are the mesh vertices; $A = (x_A, y_A), B = (x_B, y_B), C = (x_C, y_C)$; and α, β and γ are the weights of the three vertices, which conform to the following equation:

$$\alpha + \beta + \gamma = 1 \tag{2}$$

Then, Equation (1) can be converted to

$$\begin{bmatrix} x - x_A \\ y - y_A \end{bmatrix} = \beta \begin{bmatrix} x_B - x_A \\ y_B - y_A \end{bmatrix} + \gamma \begin{bmatrix} x_C - x_A \\ y_C - y_A \end{bmatrix} \tag{3}$$

Combining Equations (2) and (3), the weights of three vertices can be computed as in (4)–(6).

$$\beta = \frac{(x - x_A)(y_C - y_A) - (y - y_A)(x_C - x_A)}{(x_B - x_A)(y_C - y_A) - (y_B - y_A)(x_C - x_A)} \tag{4}$$

$$\gamma = \frac{(x - x_A)(y_B - y_A) - (y - y_A)(x_B - x_A)}{(x_C - x_A)(y_B - y_A) - (y_C - y_A)(x_B - x_A)} \tag{5}$$

$$\alpha = 1 - \beta - \gamma \tag{6}$$

$Mesh = \{A, B, C\}$ is transferred to $Mesh' = \{A', B', C'\}$ in the UV coordinate system, so the mapping of target point P is defined as $P' = \alpha A' + \beta B' + \gamma C'$. The same method can be used for the mapping of the source point, Q.

This coarse positioning step essentially decreases the variation magnitude of facial point sets between two different expressions by normalizing them all to the texture space. As shown in Figure 4, after the transformation from the image space to the texture space, each target point and source point on the left side of the figure are coarsely positioned into a fixed mesh on the right. The connecting lines between the red and blue points represent the movement trend of the set of points. After coarse positioning, the motion intensity of the point set is weakened. We show ablation studies in Section 3.

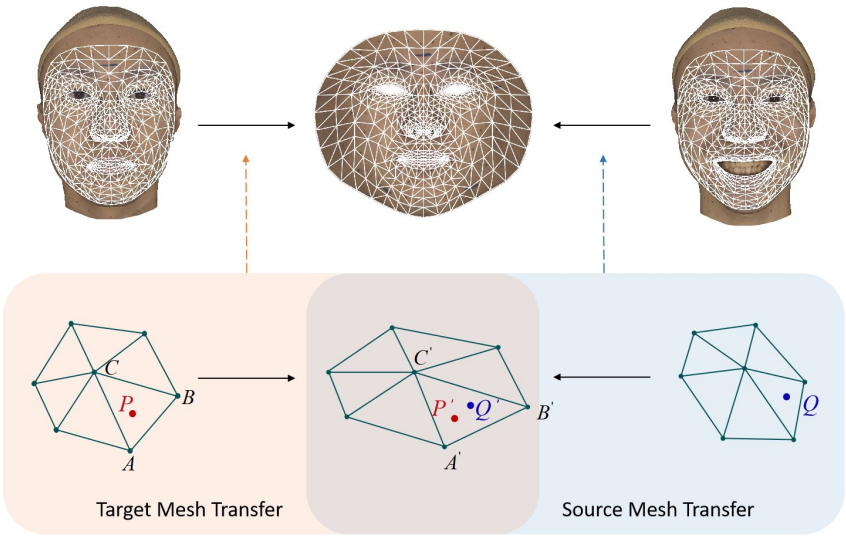

Figure 3. Transfer of marker and mesh.

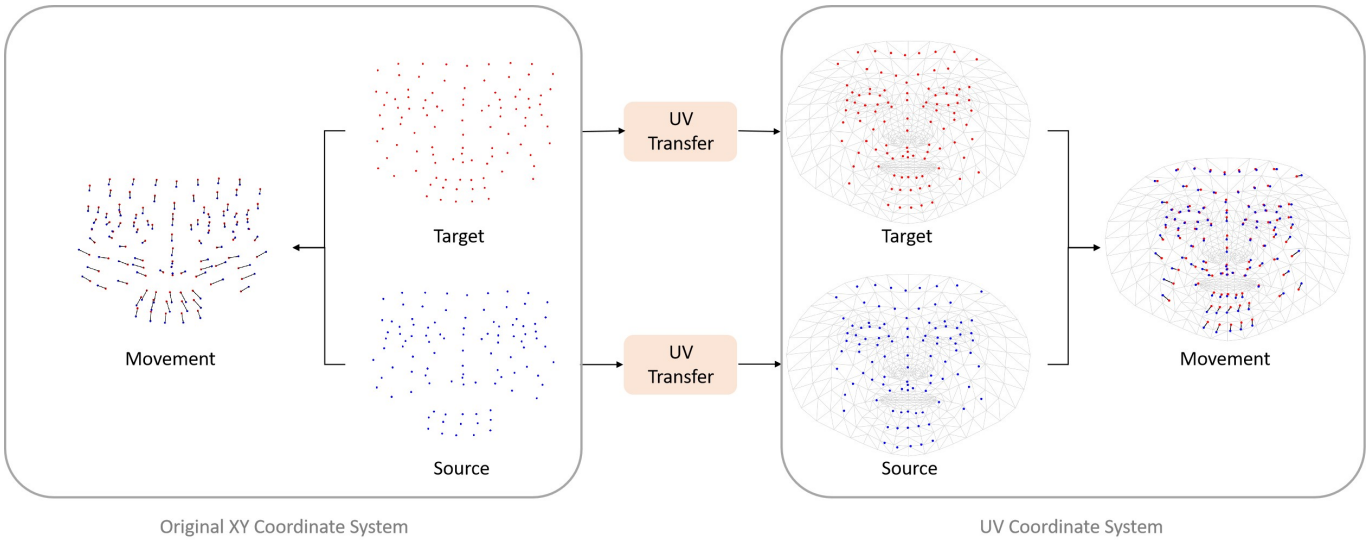

Figure 4. Movement comparison in different coordinate systems.

Moreover, owing to the invariant structure of the mesh in the texture space, there is no need for preprocessing steps such as rotation and size normalization for the extracted point set; so, this method is more suitable for complex and various situations, including extreme expression changes or faces of different scales.

2.3. Fine Matching

This subsection introduces the traditional SC descriptor [33] and analyzes its strengths and weaknesses in describing the local structure of the point set. On this basis, we propose an improved method for face models to solve the mismatch problem in most cases.

2.3.1. Shape Context Descriptor

The traditional SC is used to represent the shape features and describes a set of n 2D points as a set of n histograms $S = \{S_n \mid n = 1, 2, \ldots, N\}$. More precisely, Figure 5 shows the SC feature diagram of point p in a face point set with 78 markers. The polar coordinate system is divided into $R \times \Theta$ bins, where $R = 5$ denotes the number of radial divisions and $\Theta = 12$ is the tangential direction. The bins are uniform in log-polar space, making the descriptor more sensitive to positions of nearby points than to those of more distant points [33].

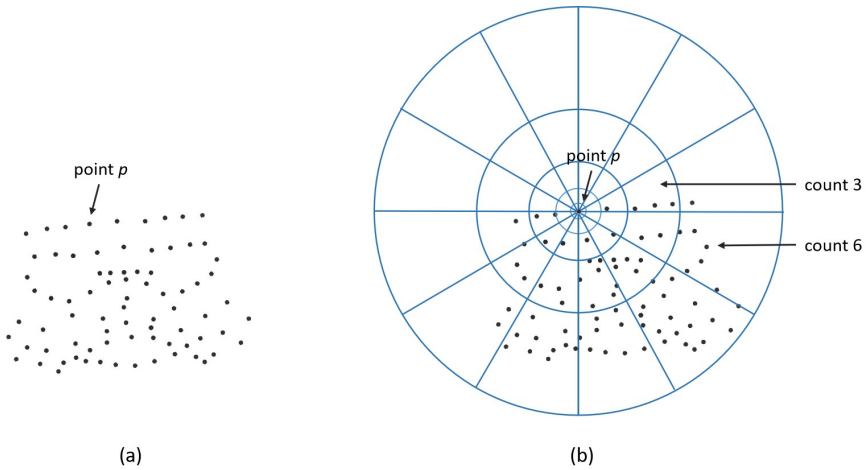

Figure 5. A scheme of representation of point sets using SC. (**a**) A point set of 78 points. (**b**) The SC feature diagram of point p.

For each point $p_n(n = 1, \ldots, N)$ of shape P, its log-polar histogram $S_n \in \mathbf{R}^{R \times \Theta}$ is computed using the relative coordinates of the remaining $n - 1$ points. The similarity between two points (p_n of P and q_m of Q) can be measured using the χ^2 statistics.

$$d_{sc}^{nm} = d_{sc}(S_n(P), S_m(Q)) = \frac{1}{2} \sum_{r=1}^{R} \sum_{\theta=1}^{\Theta} \frac{\left(s_{r\theta}^{n}(P) - s_{r\theta}^{m}(Q)\right)^2}{s_{r\theta}^{n}(P) + s_{r\theta}^{m}(Q)} \tag{7}$$

where $s_{r\theta}^{n}(P)$ is the element in the rth row and θth column of matrix $S_n(P)$, equal to the number of points in this bin. The cost matrix between the two point sets is computed as $D_{sc} = \{d_{sc}^{nm}\}_{n=1,m=1}^{N,M}$ to describe the similarity between each pair of points of the two sets. Then, the point-to-point correspondences can be obtained by minimizing the sum of the matching costs, which is obtained as

$$Cost(\Pi) = \sum_{n} d_{sc}^{n,\pi(n)} \tag{8}$$

where Π is the computed permutation of shape Q corresponding to shape P and $\pi(n)$ is the nth entry of Π.

The traditional SC features are usually used to describe the distribution of contour sampling points with high shape recognition accuracy [33]. However, due to the complexity of the actual object shape, matching based on the SC descriptor is not ideal in some cases, and feature representation is not accurate enough. The problem is mainly manifested in two aspects: First, the SC descriptor does not have the rotation-invariant property, that is, it cannot effectively match under conditions of arbitrary angular poses and is sensitive to angular changes. To address this issue, we elaborated and improved the coarse positioning stage. Second, uneven movement between sampling points and the local structural similarity between adjacent points may lead to a decline in discrimination and affect the matching results. We propose an improvement plan in the next subsection.

2.3.2. Additional Correction

The SC feature of markers may change when markers move with face deformation. Given two markers with similar local structures, using a single SC feature descriptor does not enable them to be distinguished from each other effectively, resulting in regional cross-matching. In the fine matching stage, we propose a shape context descriptor with added correction (SC-AC) to reduce mismatch while matching markers of different expressions.

The target points of the neutral expression and the source points of other expressions are represented by red and blue points, respectively, in Figure 6. After feature description, performed by the SC descriptor, the elements of the cost matrix can be obtained with Equation (7). Then, we can calculate the minimum cost based on the Munkres algorithm to obtain the opti-

mal solution [38]. We connect the estimated pairs with lines in the same figure (represented by black connecting lines in Figure 6). If a cross occurs, it indicates that there may be mismatch (denoted by cyan crossing lines in Figure 6).

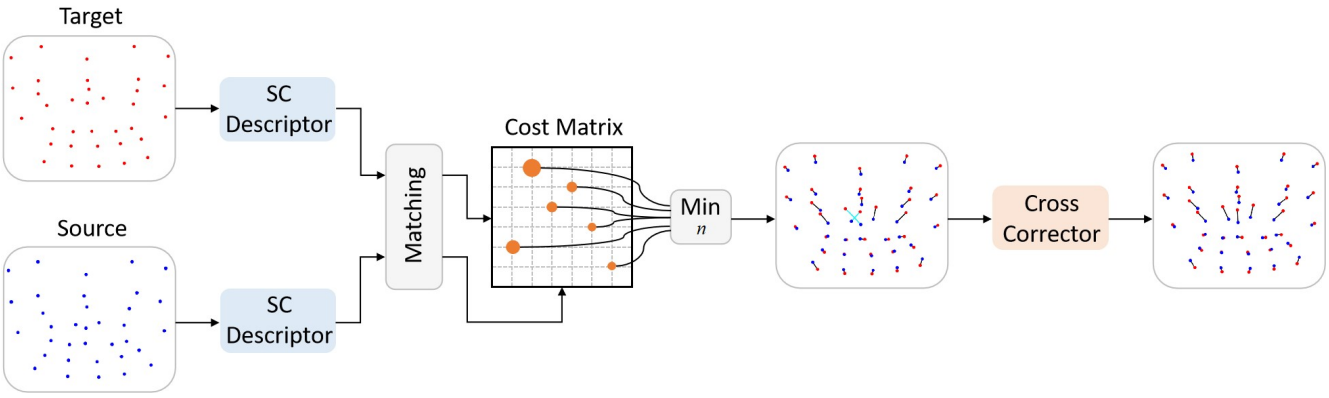

Figure 6. Overview of the fine matching process.

Cross-mismatch may be caused by uneven movement of facial sampling points. Moreover, adjacent points have a similar local structure. Especially after the face has undergone movement, the feature difference in adjacent points may be reduced, resulting in discrimination error. Cross-mismatch disrupts the mesh topology we have set in advance. Additional crossing detection can alert us that there is topological irrationality somewhere.

On this occasion, the SC feature descriptor ceases to be effective. We reintroduce the European distance feature and only correct this possible mismatch with the cross corrector, which is based on the closest principle. Considering various factors, the SC-AC method combines SC and European distance features to decrease regional cross-mismatch well. The comparison can be found in Section 3.

3. Experiments

3.1. Experimental Conditions

In this subsection, we first describe the dataset building process, including the collection plan, data structure, and the number of samples, and then give the name and version of the software used to implement the algorithm.

3.1.1. Datasets

We scanned one male subject and one female subject with a Light Stage system [39] to build the dataset. Before capturing, we drew some small black markers on the subjects' faces as a reference for the subsequent experiment. The markers were evenly distributed in different parts of the face. Subjects were asked to keep the back of the head and neck as still as possible, so that we could obtain complete pictures of expressive faces from a unified viewpoint. Each subject presented a neutral face and 25 expressions [40] (as shown in Figure 7), including some extreme expressions (e.g., screaming or crying), asymmetrical deformations (e.g., mouth to the left or right) and subtle motions (e.g., lip tightening or eye blinking). To simplify the process, after scanning, we rendered the scan model with a fixed frontal viewpoint to obtain 2D images. The following matching experiments were based only on the single view of expressive images, which covered as much of the facial area as possible. Each capture of an expression contained a set of 100 facial markers, among which those from neutral faces were set as target points and those from other expressions were set as source points. In conclusion, we prepared 50 sets of points as data for the following experiments.

3.1.2. Implementation Details

Our algorithms were implemented using Python (version 3.7.11), Mediapipe (version 0.8.9.1), Numpy (version 1.21.5) and Opencv (version 4.5.1.48). According to Section 2.3, $R = 5$ and $\Theta = 12$ were empirically determined to describe the bins of the SC descriptor. The value of $N = 100$, which was decided according to experience, denoted the number of target points. For possible occlusion during face movements, the value of M, which denoted the number of source points, was set to 100, 95, 90, 85, 80 and 75 in the anti-interference test, respectively. The SC-UV and SC-UV-AC methods were implemented using MediaPipe Attention Mesh with 468 landmarks and 849 triangles.

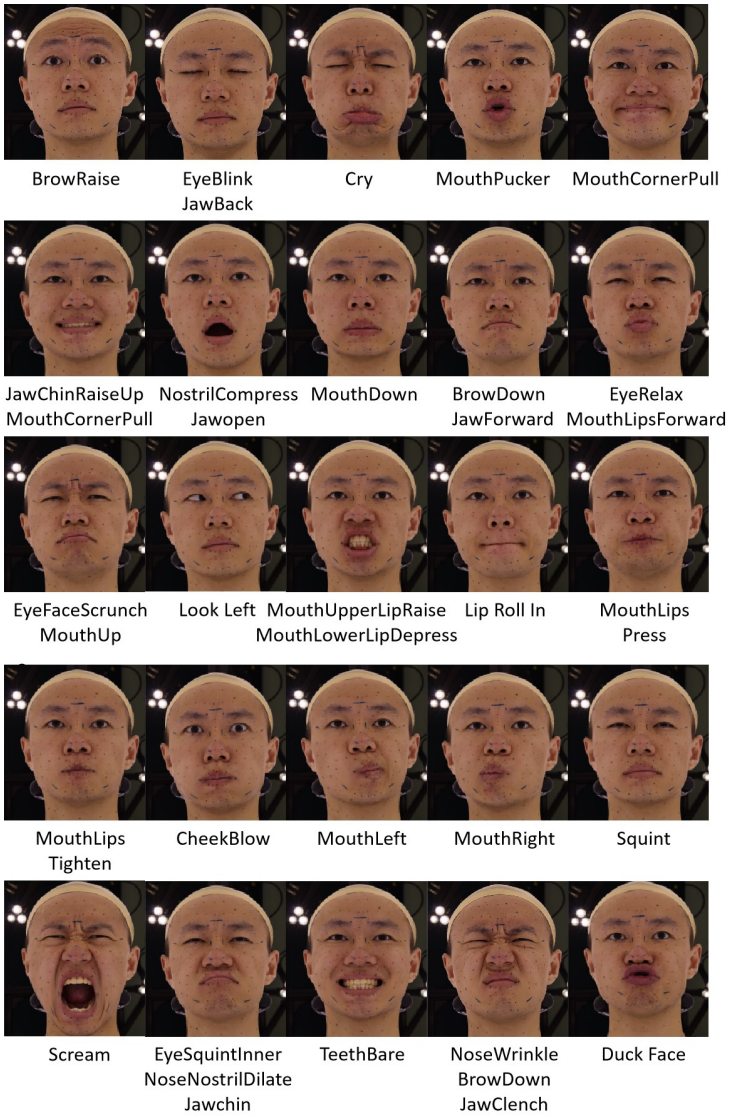

Figure 7. The 25 expressions captured.

3.2. Experimental Results

This subsection mainly introduces the experiments designed to evaluate the performance of our approach, including ablation studies, comparative experiments with other methods and tests for matching precision with interference from outliers. All experimental results are discussed.

3.2.1. Ablation Study

In order to verify the effectiveness of the SC-UV method proposed in the coarse positioning stage for face deformations, we recorded the Euclidean distance between 100 pairs of

points for each expression in our dataset relative to the neutral expression (being relaxed). After that, we selected the top 10 expressions based on the average of the distances to study the effect of the coarse positioning step on expressions with larger movements. As the width of the eyes changes very little when the face moves, we normalized the distance recorded according to the width of eye contour covered by the mesh. The average and standard deviation are shown in Figure 8, representing the moving degree of facial points when the neutral expression changes to others.

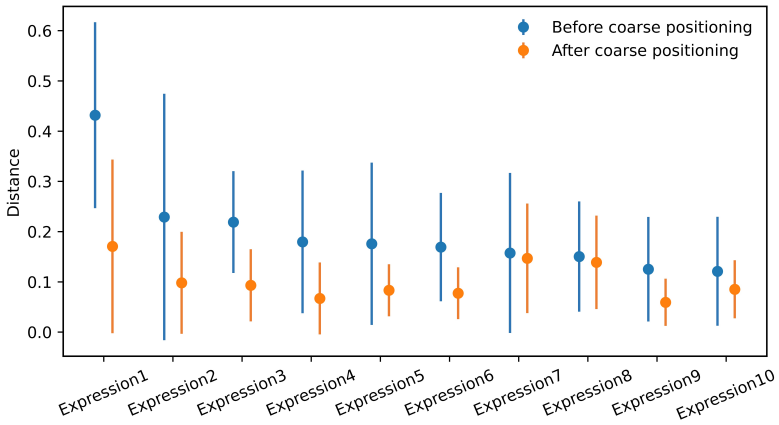

Figure 8. The distances between point pairs of different expressions. The values represent the proportion relative to the width of eye contour.

It is worth noting that for Expression 1 (screaming), the higher average distance indicates that this expression has greater facial movement overall. The large standard deviation does not mean that the data were inaccurate, but reflects the uneven movement of different parts of the face (less regarding the eye corners and more in terms of the mouth). We also calculated the total matching cost for each expression according to Equation (8), which is shown in Figure 9, indicating the shape differences.

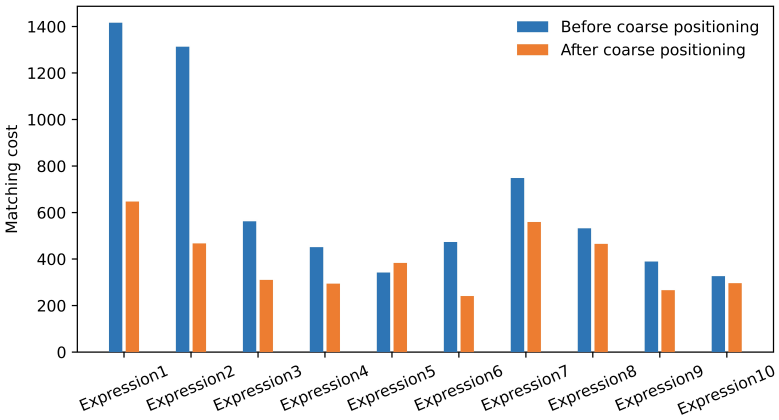

Figure 9. The matching cost for point sets of different expressions.

By comparing the data before and after coarse positioning, it could be seen that variation magnitude of facial points between two different expressions was decreased to some extent. It also means that coarse positioning could make the distribution of points on a face more standardized, which is more conducive to non-rigid registration based on the SC descriptor. In addition, the method worked better for point sets with larger face deformation, such as Expression 1 and Expression 2.

We also validated the availability of the SC-AC method proposed in the fine matching stage with the same dataset. Some cases are shown in Figure 10 to illustrate that the addi-

tional correction contributed to the results. To be specific, the above three cases represent the optimal solution for matching based on traditional SC features using the Munkres algorithm [38]. There are several regional crosses, highlighted in cyan in Figure 10. The results of re-checking and matching based on additional correction are shown in the last row, where all the estimated correspondences are in accordance with the rules of facial movement and structural features.

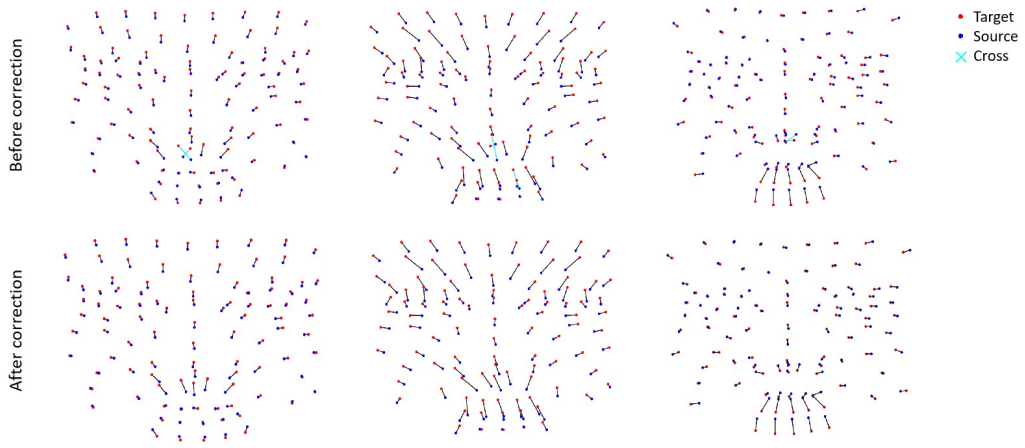

Figure 10. Cases for additional correction.

We further verified the effectiveness of these methods respectively by measuring the correspondence accuracy, as shown in Table 1. The accuracy was calculated according to the equation below:

$$\text{Accuracy} = \frac{\text{correctly matched pairs}}{\text{all input pairs}} \times 100\% \tag{9}$$

Table 1. Results of the ablation study.

Method	Accuracy % ↑
SC	94.60
SC-AC	96.70
SC-UV	98.80
SC-UV-AC	**99.80**

3.2.2. Comparison with Other Algorithms

We evaluated the performance of our SC-UV-AC method combined with the Munkres algorithm on the homemade face dataset by comparing it with previous methods. We tested the method on 25 point sets for each subject, and each set contained 100 markers. The mean correspondence accuracy for both subjects was used to determine the final matching score of the method. The comparison results are shown in Table 2, where the last row represents our method.

Table 2. Quantitative comparison of accuracy with existing methods.

Method	Accuracy of Subject 1 % ↑	Accuracy of Subject 2 % ↑	Mean Accuracy % ↑
ICP	84.60	86.76	85.68
CPD	92.84	97.24	95.04
TPS-RPM	95.52	96.12	95.82
PR-GLS	98.56	97.96	98.26
SC+Munkres	97.68	97.31	97.50
SC-UV-AC+Munkres	**99.92**	**99.85**	**99.89**

Overall, the ICP method did not work well on the self-built point sets for different subjects, which indicates that only using the Euclidean distance feature to describe the shape has a poor effect, especially when there is extreme deformation. Other methods that use a single SC descriptor to estimate the correspondence between points also have limitations. Our method combines SC and Euclidean distance features and showed superior performance over the other five algorithms.

In addition, the current research on non-rigid point cloud matching algorithms is generally based on the following two iteration steps: The first step is to evaluate the correspondence between the source point set and the target point set. The next step is to optimally update the equations of spatial transformation for the source point set. The core idea is to gradually adjust the initial position of the source point set during the iterative process, so that it can become more similar to the target point set, making it easier to evaluate the correspondence between the point sets and ultimately achieving accurate alignment between the two point sets. Algorithms based on the iterative idea, such as ICP, CPD, TPS-RPM and PR-GLS mentioned above, all showed low accuracy in the initial stage of matching, and it took them multiple rounds of position updating and recalculation to achieve the final result. Our method, on the other hand, achieved over 99% accuracy after only the first round of computation, eliminating the need for multiple iterations to save on running time. This is where our method differs from other methods.

3.2.3. Test of Anti-Interference

We tested the anti-interference capability of our method in the case of outliers. If a part of the target points have no true counterpart in the source point set, these points are outliers, and those with true counterparts are inliers. Outliers usually appear due to the possible occlusion of markers during face movements.

We set different outlier ratios and randomly generated outliers in the real dataset, which contained a total of 50 sets of points (25 sets for each subject). In general, we hope that the points retained by the algorithm are matched correctly as much as possible, because it is important to us to ensure that the estimated matching correspondence is accurate when using it as reference information for practical applications such as 3D model re-topology.

Therefore, the precision rate was used in quantitative analysis to evaluate the performance of the algorithm and was calculated according to the equation below:

$$\text{Precision} = \frac{\text{correct pairs among those retained}}{\text{all estimated pairs retained}} \times 100\% \tag{10}$$

Figure 11 shows the results of our method with different outlier ratios. We repeated the test three times in each experimental condition to take the average as the final score. For different outlier ratios, the precision rate of our method was improved over the original SC matching algorithm, and it achieved more than 93% even with 25% of outliers, demonstrating more accurate and stable matching performance.

Some cases with outliers that preliminarily simulate occlusion in face deformation are shown in Figure 12. The inputs are the point sets of a neutral expression and five other expressions, which have five different outlier ratios. The outputs are the results of the calculated correspondences between them. The black lines connect the pairs of red and blue dots. The cyan dots represent the outliers estimated by the algorithm.

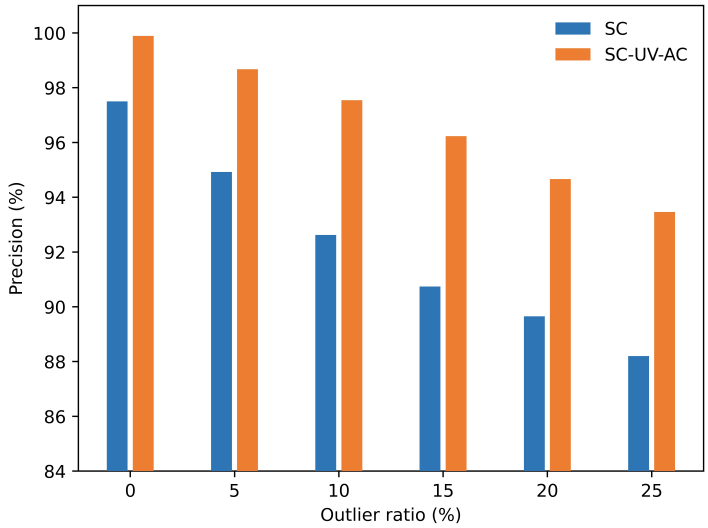

Figure 11. Results of anti-interference test.

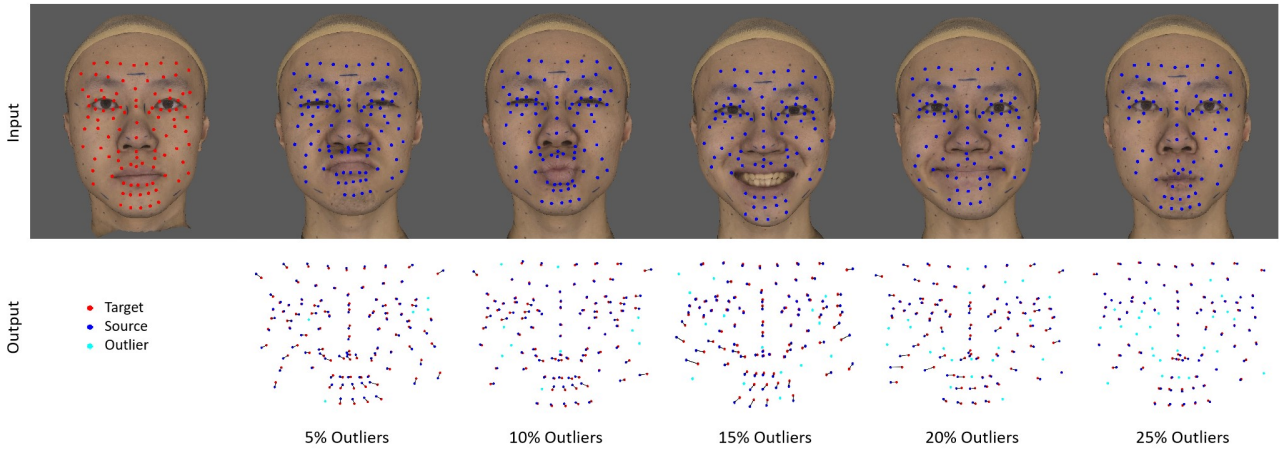

Figure 12. Matching cases for point sets with outliers.

4. Conclusions

In this paper, we propose a new elastic matching method named SC-UV-AC. Firstly, it introduces a step of coarse positioning based on the invariant structure of the texture space, which decreases the variation magnitude of facial markers between two different expressions. Secondly, it makes a modification to the traditional SC feature with additional correction to effectively reduce regional cross-mismatch.

The advantage of this method is that it does not require preprocessing steps such as face rotation and size normalization. Moreover, it can extract and match the unique facial markers of each character without the need to draw uniform markers for all characters. The estimated correspondences of markers provide a useful reference for modelers to build the topology of face models, reducing workload when creating digital avatars.

There are also some shortcomings. The performance of our method is limited by the detection region and accuracy of Attention Mesh. Moreover, in order to simplify the algorithm, we matched markers of images with different expressions in 2D and then projected them onto the 3D model, instead of matching them in real 3D space.

Experiments were carried out on a self-built face dataset with five other popular methods. For face deformations, our method achieved 99.89% accuracy. In addition, ablation studies and anti-interference tests were conducted to verify the effectiveness of our method. Higher accuracy and robustness demonstrate that our approach is more suitable for situations where the face model has extreme expression changes or a part of the target points are occluded.

Our next step is to extend our method to areas outside the mesh, including the forehead, neck and ears, to break the limitations of Attention Mesh to the greatest extent possible. In addition, we would like to explore ways to ensure temporal coherency for video sequences with extreme face deformations.

Author Contributions: Conceptualization, Y.L.; methodology, Y.L. and D.W.; software, Y.L.; validation, D.W. and Y.L.; formal analysis, Y.L.; investigation, Y.L.; resources, D.W.; data curation, Y.L., D.W. and J.C.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L., D.W. and J.C.; visualization, Y.L.; supervision, D.W.; project administration, Y.L.; funding acquisition, D.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (No.2022YFF0902303) and the Beijing Municipal Science & Technology Commission and Administrative Commission of Zhongguancun Science Park under Grant Z221100007722002.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data can be made available upon request from the authors.

Acknowledgments: We would like to thank Beijing Engineering Research Center of Mixed Reality and Advanced Display for providing experimental materials and instruments as well as the associate editor and the anonymous reviewers for their helpful feedback.

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

References

1. Fang, Z.; Cai, L.; Wang, G. MetaHuman Creator The starting point of the metaverse. In Proceedings of the 2021 International Symposium on Computer Technology and Information Science (ISCTIS), Guilin, China, 4–6 June 2021; pp. 154–157.
2. Zhang, X.; Yang, D.; Yow, C.H.; Huang, L.; Wu, X.; Huang, X.; Guo, J.; Zhou, S.; Cai, Y. Metaverse for Cultural Heritages. *Electronics* **2022**, *11*, 3730. [CrossRef]
3. Feng, Y.; Feng, H.; Black, M.J.; Bolkart, T. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Trans. Graph.* **2021**, *40*, 1–13. [CrossRef]
4. Riviere, J.; Gotardo, P.; Bradley, D.; Ghosh, A.; Beeler, T. Single-shot high-quality facial geometry and skin appearance capture. *ACM Trans. Graph.* **2020**, *39*, 1–12. [CrossRef]
5. Fang, Z.; Cai, L.; Wang, G. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Berlin, Germany; pp. 53–70.
6. Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; Tong, X. Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 16–20 June 2019; IEEE: Piscataway, NJ, USA, 2019.
7. Nagano, K.; Seo, J.; Xing, J.; Wei, L.; Li, Z.; Saito, S.; Agarwal, A.; Fursund, J.; Li, H.; Roberts, R.; et al. paGAN: Real-time avatars using dynamic textures. *ACM Trans. Graph.* **2018**, *37*, 1–12. [CrossRef]
8. Yoon, J.S.; Shiratori, T.; Yu, S.I.; Park, H.S. Self-supervised adaptation of high-fidelity face models for monocular performance tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; IEEE: Piscataway, NJ, USA; pp. 4601–4609.
9. Zollhöfer, M.; Thies, J.; Garrido, P.; Bradley, D.; Beeler, T.; Pérez, P.; Stamminger, M.; Nießner, M.; Theobalt, C. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Comput. Graph. Forum* **2018**, *37*, 523–550. [CrossRef]
10. Wu, F.; Bao, L.; Chen, Y.; Ling, Y.; Song, Y.; Li, S.; Ngan, K.N.; Liu, W. Mvf-net: Multi-view 3d face morphable model regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; IEEE: Piscataway, NJ, USA; pp. 959–968.
11. Gotardo, P.; Riviere, J.; Bradley, D.; Ghosh, A.; Beeler, T. Practical Dynamic Facial Appearance Modeling and Acquisition. *ACM Trans. Graph.* **2018**, *37*, 1–13. [CrossRef]

12. Lombardi, S.; Simon, T.; Saragih, J.; Schwartz, G.; Lehrmann, A.; Sheikh, Y. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.* **2019**, *38*, 1–14. [CrossRef]
13. Dou, P.; Kakadiaris, I.A. Multi-view 3D face reconstruction with deep recurrent neural networks. *Image Vis. Comput.* **2018**, *80*, 80–91. [CrossRef]
14. Liu, F.; Zhu, R.; Zeng, D.; Zhao, Q.; Liu, X. Disentangling features in 3D face shapes for joint face reconstruction and recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Piscataway, NJ, USA; pp. 5216–5225.
15. Tewari, A.; Bernard, F.; Garrido, P.; Bharaj, G.; Elgharib, M.; Seidel, H.P.; Pérez, P.; Zollhofer, M.; Theobalt, C. Fml: Face model learning from videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; IEEE: Piscataway, NJ, USA; pp. 10812–10822.
16. Zhou, Y.; Deng, J.; Kotsia, I.; Zafeiriou, S. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; IEEE: Piscataway, NJ, USA; pp. 1097–1106.
17. Guo, J.; Zhu, X.; Yang, Y.; Yang, F.; Lei, Z.; Li, S.Z. Towards fast, accurate and stable 3d dense face alignment. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Berlin, Germany; pp. 152–168.
18. Wood, E.; Baltrušaitis, T.; Hewitt, C.; Johnson, M.; Shen, J.; Milosavljević, N.; Wilde, D.; Garbin, S.; Sharp, T.; Stojiljković, I.; et al. 3d face reconstruction with dense landmarks. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Berlin, Germany; pp. 160–177.
19. Egger, B.; Smith, W.A.; Tewari, A.; Wuhrer, S.; Zollhoefer, M.; Beeler, T.; Bernard, F.; Bolkart, T.; Kortylewski, A.; Romdhani, S.; et al. 3d morphable face models—Past, present, and future. *ACM Trans. Graph.* **2020**, *39*, 1–38. [CrossRef]
20. Li, T.; Liu, S.; Bolkart, T.; Liu, J.; Li, H.; Zhao, Y. Topologically Consistent Multi-View Face Inference Using Volumetric Sampling. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA; pp. 3824–3834.
21. Li, J.; Kuang, Z.; Zhao, Y.; He, M.; Bladin, K.; Li, H. Dynamic facial asset and rig generation from a single scan. *ACM Trans. Graph.* **2020**, *39*, 1–18. [CrossRef]
22. Seol, Y.; Ma, W.C.; Lewis, J.P. Creating an actor-specific facial rig from performance capture. In Proceedings of the 2016 Symposium on Digital Production (DigiPro), Anaheim, CA, USA, 23 July 2016; ACM: New York, NY, USA; pp. 13–17.
23. Taylor, J.; Bordeaux, L.; Cashman, T.; Corish, B.; Keskin, C.; Sharp, T.; Soto, E.; Sweeney, D.; Valentin, J.; Luff, B.; et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Trans. Graph.* **2016**, *35*, 1–12. [CrossRef]
24. Besl, P.J.; McKay, N.D. A method for registration of 3-D shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 239–256. [CrossRef]
25. Chui, H.; Rangarajan, A. A new point matching algorithm for non-rigid registration. *Comput. Vis. Image Underst.* **2003**, *89*, 114–141. [CrossRef]
26. Sheikhbahaee, Z.; Nakajima, R.; Erben, T.; Schneider, P.; Hildebrandt, H.; Becker, A. Photometric calibration of the COMBO-17 survey with the Softassign Procrustes Matching method. *Mon. Not. R. Astron. Soc.* **2017**, *471*, 3443–3455. [CrossRef]
27. ElSayed, A.; Kongar, E.; Mahmood, A.; Sobh, T.; Boult, T. Neural Generative Models for 3D Faces with Application in 3D Texture Free Face Recognition. 2018. Available online: https://arxiv.org/pdf/1811.04358.pdf (accessed on 20 December 2022).
28. Myronenko, A.; Song, X. Point set registration: Coherent point drift. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 2262–2275. [CrossRef] [PubMed]
29. Yuille, A.L.; Grzywacz, N.M. A mathematical analysis of the motion coherence theory. *Int. J. Comput. Vis.* **1989**, *3*, 155–175. [CrossRef]
30. Zheng, Y.; Doermann, D. Robust point matching for nonrigid shapes by preserving local neighborhood structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 643–649. [CrossRef]
31. Jian, B.; Vemuri, B.C. Robust point set registration using gaussian mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 1633–1645. [CrossRef]
32. Ma, J.; Zhao, J.; Yuille, A.L. Non-rigid point set registration by preserving global and local structures. *IEEE Trans. On Image Processing* **2015**, *25*, 53–64.
33. Belongie, S.; Malik, J.; Puzicha, J. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 509–522. [CrossRef]
34. Cheng, S.; Chen, X.; Qu, H. Rapid Real-Time Collision Detection for Large-Scale Complex Scene Based on Virtual Reality. In Proceedings of the 2021 International Conference on Applications and Techniques in Cyber Intelligence (ATCI), Fuyang, China, 19–21 June 2021; Abawajy, J., Xu, Z., Atiquzzaman, M., Zhang, X., Eds.; Springer: Cham, Switzerland; pp. 605–610.
35. ElSayed, A.; Kongar, E.; Mahmood, A.; Sobh, T.; Boult, T. Attention Mesh: High-Fidelity Face Mesh Prediction in Real-Time. 2020. Available online: https://arxiv.org/pdf/2006.10962.pdf (accessed on 20 December 2022).
36. Kartynnik, Y.; Ablavatski, A.; Grishchenko, I.; Grundmann, M. Real-Time Facial Surface Geometry from Monocular Video on Mobile GPUs. 2019. Available online: https://arxiv.org/pdf/1907.06724.pdf (accessed on 20 December 2022).
37. Bazarevsky, V.; Kartynnik, Y.; Vakunov, A.; Raveendran, K.; Grundmann, M. Blazeface: Sub-Millisecond Neural Face Detection on Mobile GPUs, 2019. Available online: https://arxiv.org/pdf/1907.05047.pdf (accessed on 20 December 2022).

38. Gabrovšek, B.; Novak, T.; Povh, J.; Rupnik Poklukar, D.; Žerovnik, J. Multiple Hungarian method for k-assignment problem. *Mathematics* **2020**, *8*, 2050. [CrossRef]
39. Ghosh, A.; Fyffe, G.; Tunwattanapong, B.; Busch, J.; Yu, X.; Debevec, P. Multiview face capture using polarized spherical gradient illumination. In Proceedings of the 2011 SIGGRAPH Asia Conference, Hong Kong, China, 12–15 December 2011; ACM: New York, NY, USA; pp. 1–10.
40. Ekman, P.E.; Friesen, W.V. Facial Action Coding System: A Technique for the Measurement of Facial Actions. *Riv. Psichiatr.* **1978**, *47*, 126–138.