3D reconstruction of oral structure using oral camera

by

Moxuan Yuan

(2130026188)

A Final Year Project Thesis (COMP4004; 3 Credits)
submitted in partial fulfillment of the requirements
for the degree of

Bachelor of Science (Honours)
in
Computer Science and Technology

at
BNU-HKBU
UNITED INTERNATIONAL COLLEGE

December, 2024

DECLARATION

I hereby declare that all the work done in this Project is of my independent effort. I also certify that I have never submitted the idea and product of this Project for academic or employment credits.

_____

Moxuan Yuan

(2130026188)

Date: 12/2024

BNU-HKBU

United International College

Computer Science and Technology Program

We hereby recommend that the Project submitted by Moxuan Yuan entitled "3D reconstruction of oral structure using oral camera" be accepted in partial fulfillment of the requirements for the degree of Bachelor (Honours) of Science in Computer Science and Technology Program.

_____ _____

Date: _____ **Date: _____**

ACKNOWLEDGEMENT

I would like to express my great gratitude towards my supervisor, Prof. Hongjian Shi, who had given me invaluable advice to this project.

Contents

ABSTRACT 2
1 Introduction 3
2 System Overview 5
    2.1 Flow Chat 5
3 Camera Imaging 6
    3.1 Pinhole Model 6
    3.2 Camera Geometry 7
        3.2.1 Pixel Plane 7
        3.2.2 Projection Matrix 7
        3.2.3 World Coordinate System 8
    3.3 Camera Calibration 8
        3.3.1 Calibration Device 8
        3.3.2 Extract the Camera Intrinsic Parameters 10
4 Two-view Reconstruction 11
    4.1 Triangulation 11
    4.2 Epipolar Geometry 11
    4.3 Geometric constraints 12
        4.3.1 Essential Matrix 12
        4.3.2 Fundamental Matrix 13
    4.4 Estimate the Fundamental Matrix 13
    4.5 External Camera Parameters 14
    4.6 P3P Camera Pose 14
    4.7 Bundle Adjustment 14
5 Data Pre-processing 15
    5.1 Collect Data 15
    5.2 Image Feature Points Extraction and Matching 15
        5.2.1 Matching 15
        5.2.2 RANSAC Estimate Fundamental Matrix and Find R, T 15
6 Implementation Of 3D Reconstruction 16
    6.1 Construction of Connected Graph 16
        6.1.1 Tracks 16
        6.1.2 Graph 16
    6.2 Initial Point Cloud for Two-View Reconstruction 16
    6.3 Multi-View Reconstruction 16
7 Result 17
8 Discussion 18
9 Related Work 19
    9.1 Reconstruction of Teeth 19
    9.2 Mobile phone imaging 19
10 Conclusion 20
References 20

ABSTRACT

As oral health awareness continues to grow, dental care has become an increasingly common need. However, current dental diagnostic processes face several limitations, such as requiring fixed locations, being static in nature, and lacking effective visualization. These challenges reduce diagnostic efficiency, limit patient experience, and increase the difficulty of accurate diagnosis for dentists. To address these issues, this paper proposes a novel method for reconstructing teeth using 3D reconstruction technology. While 3D reconstruction has been applied in fields like medical imaging and virtual reality, its accessibility in dental diagnostics remains limited. Current methods often rely on expensive professional equipment, hindering widespread adoption. The study leverages smartphone photography as a convenient data source for 3D reconstruction. By capturing multiple images of teeth from different angles, we reverse the imaging process using a nonlinear multi-view triangulation method to reconstruct a 3D tooth model. To improve robustness against noise, inconsistent angles, and imperfect lighting, we incorporate an incremental reconstruction strategy that optimizes the model step-by-step. Extensive experiments demonstrate that the proposed system achieves high accuracy and usability. With only a smartphone, users can generate high-quality 3D tooth models, enabling flexible and remote dental diagnostics.

1 Introduction

As the times advance, the importance of dental healthcare in overall physical well-being is steadily increasing. Traditional medical diagnostics in dentistry often relied heavily on the practitioner's personal judgment, such as the most conventional method of manually inspecting the oral cavity without auxiliary tools or using CT scans to obtain dental radiographs. Recently, innovations such as the So prolife fluorescence caries detection device and direct digital imaging systems have emerged as key research areas. Among these, CT scanning has gained widespread application due to its convenience, efficiency, and high accuracy.

CT scanning, which records cross-sectional images of the body using X-ray beam scanning, is widely used in various medical projects because of its requirement for high precision. However, this scanning method relies heavily on specialized equipment and provides limited visualization due to its cross-sectional nature. Additionally, the reliance on X-ray beam scanning raises health concerns, as patients are increasingly worried about radiation exposure.

The application of three-dimensional (3D) reconstruction technology in dental healthcare offers a convenient solution to the limitations of traditional visualization methods. By providing a 3D restoration of dental structures, this technology enables comprehensive observation and facilitates the digital flow of information. Cone Beam Computed Tomography (CBCT) employs CT imaging for 3D reconstruction. Its principle involves the use of a low-dose X-ray beam that rotates around the object in a circular trajectory, performing digital radiography (DR). The multiple digital projections obtained through this rotational process are subsequently reconstructed in a computer to generate 3D images [1].

However, this method requires specialized and often bulky equipment, which compromises portability. Moreover, it still relies on X-ray beam technology, leaving issues related to convenience and health risks unresolved. Similarly, CAD/CAM-based intraoral scanners can be used for dental imaging but are dependent on expensive, dedicated hardware and still pose risks related to radiation exposure.

This paper proposes a vision-based dental imaging system that utilizes image processing techniques to achieve accurate imaging. Compared to traditional CT scanning and intraoral scanners, vision-based methods are more intuitive and freer from radiation risks, making them a safer alternative for data collection. Such a system has the potential to use flexible equipment to capture surface data of teeth, including their shapes, offering a promising avenue for advancements in dental imaging technology.

In order to collect data on the surface of the teeth, we used the most convenient tool, a mobile phone. Everyone has a mobile phone, and nowadays they have excellent photographic capabilities, so such a convenient and handy tool became the focus of our study. It was a challenge because taking pictures of teeth with a mobile phone is a difficult task; the cheeks and mouth would block the complete tooth form. So, the initial tool was an intraoral camera, but then it was found that such a tool conflicted with the convenience that it was initially striving to overcome, so how to overcome the interference of taking pictures of teeth with a mobile phone was mostly the first challenge carried out. To solve this challenge, we segmented the original image using the convolutional neural network Unet [2] to obtain an image that was mostly teeth, and thus the data captured with the mobile phone had usability.

Reconstruction is a very hot topic in computer vision; stereo vision, monocular depth estimation, structured light and so on are all good methods. Ultimately, this paper uses the

Multi-View Reconstruction technique structure from motion, which has the advantages of good flexibility (no need to be scene-specific), low cost (a 2D image taken with a mobile phone is sufficient), and high efficiency (the process is systematic).

The rest of this article is structured as follows. Section 2 provides an overview of the system design, Section 3 details camera imaging, and Section 4 describes two-view reconstruction, part 5 lists the Data pre-processing, part 6 describes the implementation of 3D reconstruction, Part 7 shows the experimental results, and Part 8 discusses the limitations and possible solutions, part 9 summarizes the related work, and Section 10 summarizes this article.

2 System Overview

The 3D dental reconstruction system consists of three parts. The first part performs data preprocessing, taking an image set as input to obtain feature points after geometric validation. The second part utilizes the fundamental matrix obtained from the first part to calculate the camera intrinsic parameters. The third part conducts incremental reconstruction.

In the following chapters, the third and fourth sections first provide the technical foundations for computer-based 3D reconstruction, followed by the complete process of 3D dental reconstruction.

2.1 Flow Chat

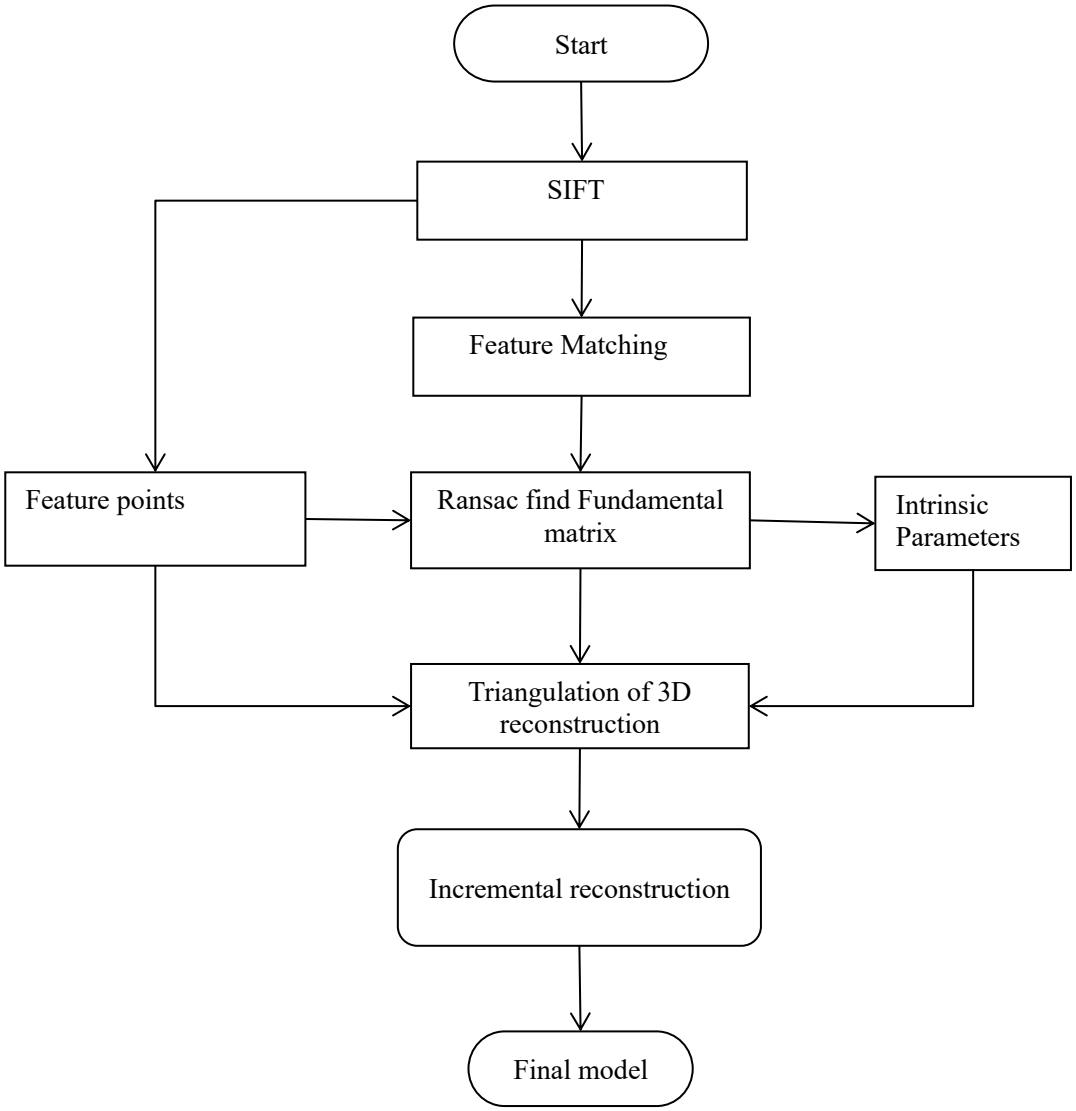

Figure 1: Flow chat

3 Camera Imaging

3.1 Pinhole Model

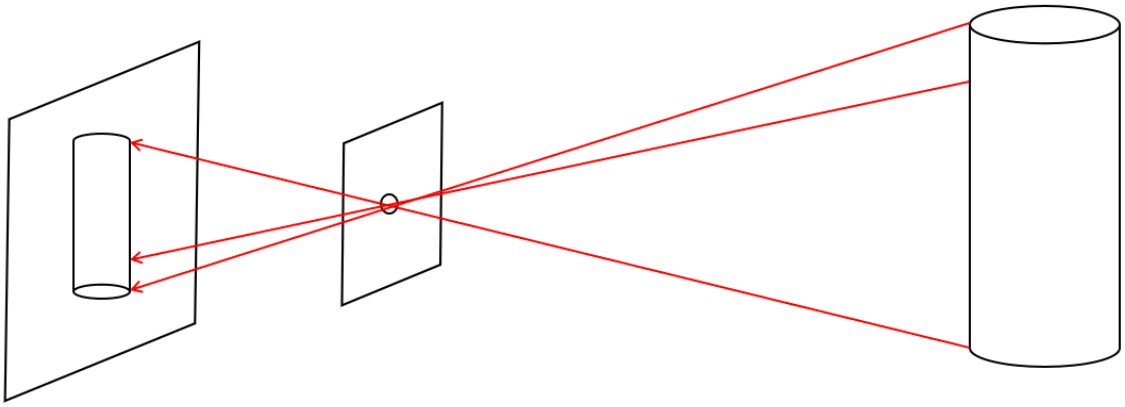

Figure 2: Pinhole camera

The first thing we need to know in order to reconstruct the original 3D shape from the image we took is how the 3D object is projected onto the 2D image. From the principle of pinhole imaging, we can know that the point on the object travels in a straight line through the light and projects to another plane through the pinhole. So, we have the pinhole camera in Figure 2; Take Pinhole as Origin O, we can obtain a camera coordinate system with axes defined by (k, j, i), The point extending in the k direction to the image plane is denoted as C', This point serves as the origin of the image plane coordinates. As shown in Figure 3.

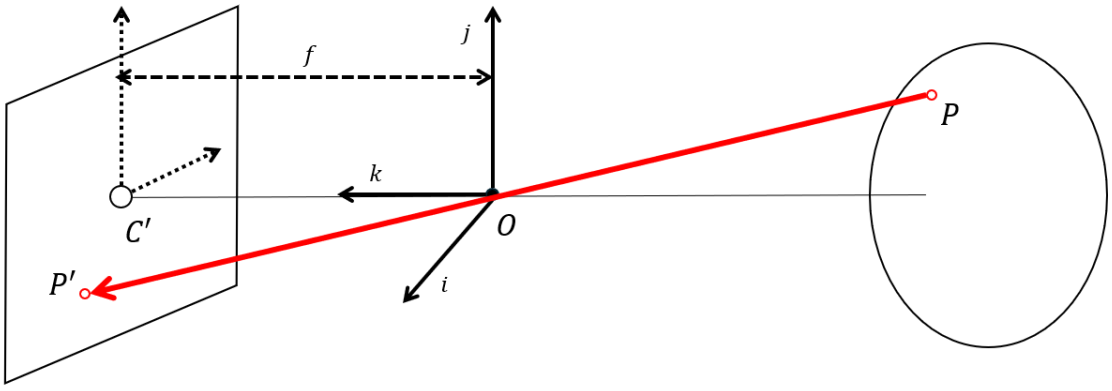

Figure 3: Camera coordinate system

We can assume that there exists a point P in reality, which, after being projected by a pinhole camera, appears on the image plane as P'. We define the coordinates of point P in the camera coordinate system as $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ and the coordinates of point P' in the image plane

coordinate system as $\begin{bmatrix} x' \\ y' \end{bmatrix}$. The distance from the camera coordinate system origin O to the image plane origin C' is f, which is the focal length. Therefore, using the method of similar triangles we can have:$\frac{y'}{f} = \frac{y}{z}$, so it derived that the y'-coordinate of P' is $y' = f\frac{y}{z}$. Similarly, we can derive that the x'-coordinate is $x' = f\frac{x}{z}$. Thus, we have established the transformation relationship for the projection of point P in the camera coordinate system to point P' on the image plane, given by $P = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \rightarrow P' = \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} f\frac{x}{z} \\ f\frac{y}{z} \end{bmatrix}$.

3.2 Camera Geometry

3.2.1 Pixel Plane

Now, with the relationship between the shooting point and the camera coordinates known, we can determine the relationship between the projected point on the image and the center of the image. However, in reality, imaging is done in pixels rather than in meters, which represent the distances in the camera coordinate system. Additionally, the origin of the image is at the lower left corner of the digital image. Therefore, we need a step to convert the object of study from the image plane to the pixel plane.

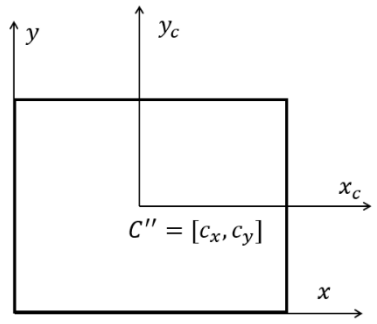

Figure 4 Pixel plane

We assume that the origin of the transformed image plane(Figure 4), in a coordinate system where the lower left corner is the origin, is denoted as C'', with coordinates $C'' = [c_x, c_y]$. According to the translation rule of coordinate systems, the coordinates of point P' are $(f\frac{x}{z} + c_x, f\frac{y}{z} + c_y)$.

Next, we need to convert the units from meters to pixels. We define two parameters, k and l, as the conversion factors for the horizontal and vertical coordinates, respectively, with units of pixels/m. The values of k and l depend on the nature of the camera's imaging components. The reason for defining two different parameters is that the imaging quality along the x-direction and y-direction often varies due to manufacturing limitations. Thus, the coordinates of the point P', after unit conversion, are $\left(fk\frac{x}{z} + c_x, fl\frac{y}{z} + c_y\right)$. Since f, k, and l are all fixed parameters, this equation can also be expressed as $\left(\alpha\frac{x}{z} + c_x, \beta\frac{y}{z} + c_y\right)$.

3.2.2 Projection Matrix

Since $\frac{\alpha}{z}$ and $\frac{\beta}{z}$ are not determined real numbers, the transformation from point P to point P' is a nonlinear transformation. To facilitate calculations, we need to introduce the concepts of homogeneous coordinates and Euclidean coordinates. Homogeneous coordinates represent an extended form of geometric coordinate points. For example, the homogeneous coordinates of a point $(x,\ y)$ in a 2D image transform to $\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$, while the

homogeneous coordinates of a spatial point (x, y, z) are $\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$. When solving for P', it is known that the farther the point is from the camera (the larger z is), the projection point $\left(\alpha\frac{x}{z} + c_x, \beta\frac{y}{z} + c_y\right)$ will be closer to the center of the image plane. Therefore, when converting a 3D point to a 2D point, dividing by the third parameter w is equivalent to dividing by z. Thus, the correspondence between 3D point coordinates in homogeneous space and points on a 2D plane is $\begin{bmatrix} x \\ y \\ \omega \end{bmatrix} \Rightarrow (\frac{x}{\omega}, \frac{y}{\omega})$.

From this, we can derive the homogeneous coordinates of P' and P as : $P'_h = \begin{bmatrix} \alpha x + c_x z \\ \beta y + c_y z \\ z \end{bmatrix} = \begin{bmatrix} \alpha & 0 & c_x & 0 \\ 0 & \beta & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$ and $P_h = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$. Therefore, $P' = MP$ where $M = \begin{bmatrix} \alpha & 0 & c_x & 0 \\ 0 & \beta & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \alpha & 0 & c_x \\ 0 & \beta & c_y \\ 0 & 0 & 1 \end{bmatrix} [I \quad 0]$. (In the following sections, P' will be used to represent P'_h, and P will be used to represent P_h.). It is important to note that in reality, the pixels captured by the camera are not perfect squares; they are typically a parallelogram with an angle of θ degrees of skew. However, in this experiment, we are using smartphone photography, which involves computational photography algorithms. Therefore, we will neglect this minor error and define θ as 90°.

3.2.3 World Coordinate System

In the previous sections, we have been using the coordinates of point P in the camera coordinate system for calculations. However, during capturing, we often do not know the distance of that point from the camera. Therefore, we need to introduce a world coordinate system as an objective scale, which serves as the foundational coordinate system for 3D reconstruction.

Thus, we need to express the coordinates of point P in the world coordinate system in terms of the camera coordinate system. After applying rotation (R) and translation (T), we can determine that $P = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} P_w$. Therefore, $P' = K[I \quad 0]P = K[I \quad 0]\begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} P_w = K[R \quad T]P_w = MP_w$.

3.3 Camera Calibration

3.3.1 Calibration Device

Now we have established the mapping relationship from the 3D world to 2D pixels, but the specific values are unknown. Therefore, solving for the camera intrinsic and extrinsic parameter matrices, or camera calibration, is an important issue in studying camera geometry.

First, we need to identify the target that needs to be calibrated, which is the projection matrix M. The intrinsic matrix K ($\begin{bmatrix} \alpha & 0 & c_x \\ 0 & \beta & c_y \\ 0 & 0 & 1 \end{bmatrix}$) has four parameters, while the extrinsic matrix $R\ and\ T$ (with three parameters for each of the x, y, z directions) has six

parameters, totaling ten parameters.

To solve for the camera intrinsic parameters, we set up a calibration device (Figure 5) composed of three planes covered with a grid. After capturing images, we can select any point P_i located on the calibration device and find the corresponding pixel coordinates $p_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix}$ in the image based on the correspondence of the grid. Thus, we have the correspondence $p_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix} = MP_i$.

We can decompose M into $\begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix}$, According to the transformation between homogeneous coordinates and Euclidean coordinates, we obtain: $p_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} \frac{m_1 P_i}{m_3 P_i} \\ \frac{m_2 P_i}{m_3 P_i} \end{bmatrix}$, allowing us to formulate the equation : $m_1 P_i - m_3 P_i u_i = 0$; $m_2 P_i - m_3 P_i v_i = 0$. As mentioned earlier, there are a total of ten parameters to be solved, so a minimum of five points is required. To obtain a more robust result, we select ten points for solving. By combining these ten equations, we obtain $\begin{cases} m_1 P_1 - m_3 P_1 u_1 = 0 \\ m_2 P_1 - m_3 P_1 v_1 = 0 \\ \vdots \\ m_1 P_n - m_3 P_n u_n = 0 \\ m_2 P_n - m_3 P_n u_n = 0 \end{cases}$, treating it as $Pm = 0$ for $P \overset{\text{def}}{=} \begin{pmatrix} P_1^T & 0 & -u_1 P_1^T \\ 0 & P_1^T & -v_1 P_1^T \\ & \vdots & \\ P_n^T & 0 & -u_n P_n^T \\ 0 & P_n^T & -v_n P_n^T \end{pmatrix}_{2n \times 12}$ and $m \overset{\text{def}}{=} \begin{pmatrix} m_1^T \\ m_2^T \\ m_3^T \end{pmatrix}_{12 \times 1}$, resulting in an overdetermined homogeneous linear equation system.

At this point, we can use the least squares method to solve for m. We perform singular value decomposition on the matrix P, obtaining UDV^T, and take the last column of the V matrix as the solution for m (the right eigenvector corresponding to the smallest eigenvalue), ensuring that $\|m\| = 1$. Subsequently, we can use the obtained m to derive $m \overset{\text{def}}{=} \begin{pmatrix} m_1^T \\ m_2^T \\ m_3^T \end{pmatrix} \Rightarrow M = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix}$.

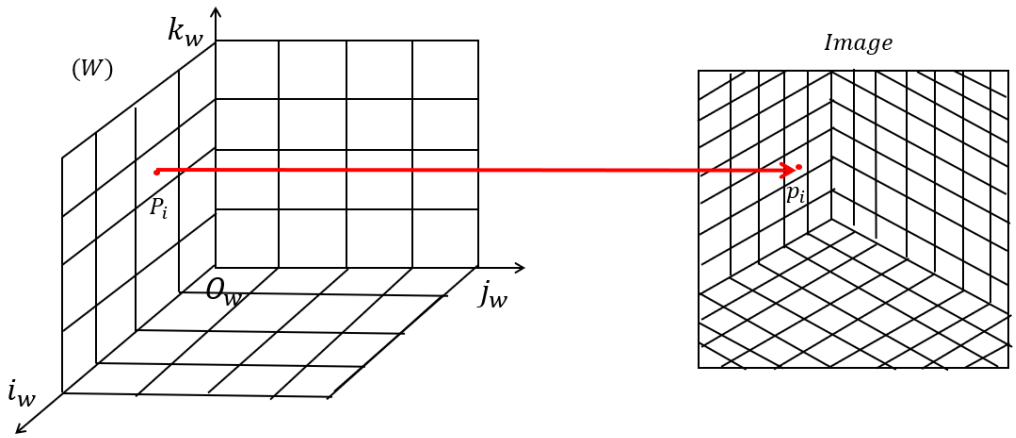

Figure 5 Calibration device

3.3.2 Extract the Camera Intrinsic Parameters

According to the definition of the projection matrix, we have $M = K[R \quad T]$, $K = \begin{bmatrix} \alpha & 0 & c_x \\ 0 & \beta & c_y \\ 0 & 0 & 1 \end{bmatrix}$, $R = \begin{bmatrix} r_1^T \\ r_2^T \\ r_3^T \end{bmatrix}$, and $T = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}$. Therefore, $\rho M = \begin{pmatrix} \alpha r_1^T + u_0 r_3^T & \alpha t_x + u_0 t_z \\ \beta r_2^T + v_0 r_3^T & \beta t_y + v_0 t_z \\ r_3^T & t_z \end{pmatrix}$. We can decompose M into $[A \quad b]$, where $A = \begin{bmatrix} a_1^T \\ a_2^T \\ a_3^T \end{bmatrix}$ and $b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$.

By calculating, we can list $\rho A = \rho \begin{pmatrix} a_1^T \\ a_2^T \\ a_3^T \end{pmatrix} = \begin{pmatrix} \alpha r_1^T + u_0 r_3^T \\ \beta r_2^T + v_0 r_3^T \\ r_3^T \end{pmatrix} = KR$. By using the dot product and cross product, we can solve for $\rho = \frac{1}{|a_3|}$(Since the calibration board may not be completely positioned in front of the camera, ρ is a positive number.), $c_x = \rho^2(a_1 \cdot a_3)$ and $c_y = \rho^2(a_2 \cdot a_3)$, $\alpha = \rho^2|a_1 \times a_3|$, $\beta = \rho^2|a_2 \times a_3|$. Thus, we have obtained all the intrinsic parameters of the camera.

4 Two-view Reconstruction

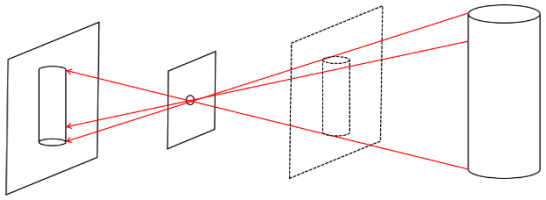

Figure 6 Back projection

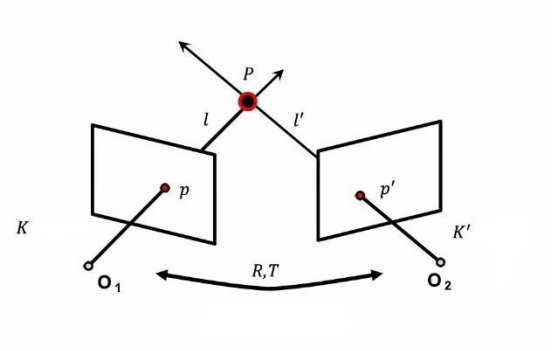

Figure 7 Triangulation

4.1 Triangulation

In the image pixel plane, we select points for back-projection (Figure 6), meaning that starting from point p on the image plane, we extend a ray backward through the origin of the camera coordinate system, indicating the possible locations of the 3D point P. The surface formed by the convergence of these points is called the projection plane. However, it is not possible to determine the depth relationships of points from a single image alone. Therefore, we introduce the important concept of triangulation in 3D reconstruction, which involves using two views to locate the position of the 3D point P.

First, we define the two views: the camera coordinate systems are denoted as O_1 and O_2, where O_2 is the coordinate system obtained by applying a rotation and translation ($R\ and\ T$) transformation to O_1. Therefore, the corresponding projection matrix for O_1 is denoted as $M = K[I \quad 0]$, and for O_2, the projection matrix is $M' = K[R \quad T]$.

In the O_1 coordinate system, the 3D point P projects onto the point p in the projection plane, while in O_2, it corresponds to point p'. The intrinsic parameters K and K' for both cameras are known. Using these known parameters, we can formulate the equations:

$$\begin{cases} p = MP = K[I \quad 0]P \\ p' = M'P = K'[R \quad T]P \end{cases} \Rightarrow \begin{cases} u = \frac{m_1 P}{m_3 P} \rightarrow m_1 P - u(m_3 P) = 0 \\ v = \frac{m_2 P}{m_3 P} \rightarrow m_2 P - v(m_3 P) = 0 \\ u' = \frac{m_1{}' P}{m_3{}' P} \rightarrow m_1{}' P - u'(m_3{}' P) = 0 \\ v' = \frac{m_2{}' P}{m_3{}' P} \rightarrow m_2' P - v'^{(m_3{}' P)} = 0 \end{cases}$$

After simplification, we obtain an overdetermined homogeneous linear equation system:

$AP = 0\ and\ A = \begin{bmatrix} um_3 - m_1 \\ vm_3 - m_2 \\ u'm_3{}' - m_1{}' \\ v'm_3' - m_2{}' \end{bmatrix}$. By applying the least squares method, we can solve for the coordinates of point P. This is the triangulation.

4.2 Epipolar Geometry

However, in practical applications, we often do not know the rotation and translation ($R\ and\ T$) between the cameras. Therefore, we need to determine the geometric relationship between the two viewpoints, which is called Epipolar geometry [2].

Connecting point P with points O_1 and O_2 forms a plane, known as the Epipolar plane. The intersection line of this plane with the projection plane is called the Epipolar line (l and l'). According to the properties of plane intersection, the corresponding points p and p' are located on lines l and l', respectively.。

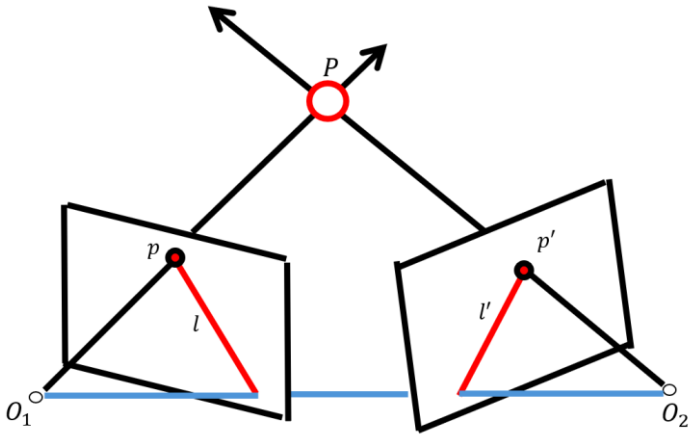

Figure 8 Epipolar Geometry

4.3 Geometric constraints

4.3.1 Essential Matrix

After understanding the relationships between points, lines, and planes within the entire model, we can derive the correspondence between two points, i.e., find a matrix that describes the Epipolar geometry relationship between two viewpoints' images. To facilitate finding this relationship, we introduce a normalized camera, where $K = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ represents the camera. The matrix that describes the Epipolar geometry under the assumption of a normalized camera is called the essential matrix.

Now, given $K = K' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, $p = (u, v)$, and $p' = (u', v')$, since the projection transformation of the normalized camera is $P' = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$, the Euclidean coordinates of a 3D point in the camera coordinate system are equivalent to the homogeneous coordinates of the image point. That is, the non-homogeneous coordinates of the spatial point p in the O_1 coordinate system are $(u, v, 1)$, and the non-homogeneous coordinates of the spatial point p' in the O_2 coordinate system are $(u', v', 1)$.

Calculations across different coordinate systems are inconvenient; therefore, we need to express the content of the O_2 coordinate system in terms of the O_1 coordinate system, that is, find p' in the O_1 coordinate system and express O_2 in O_1. Based on the rotation and translation between coordinate systems, we can obtain $p = R^T p' - R^T T$ and $O_2 = -R^T T$.

By combining the fact that the vector $O_1 p'$ and the vector $O_1 O_2$ lie in the Epipolar plane, we know that their cross product ($(R^T p' - R^T T) \times R^T T = R^T T \times R^T p'$) is perpendicular to the Epipolar plane. In other words, it is also perpendicular to the vector p. Therefore, $[R^T T \times R^T p']^T \cdot p = 0$.

After further derivation, we obtain $p'^T [T \times R] p = 0$, and thus the essential matrix E is given by $E = T \times R$.

4.3.2 Fundamental Matrix

Although the process of solving the essential matrix is smooth, it is ultimately based on the Epipolar geometry of a normalized camera. What we actually need is the fundamental matrix, which algebraically describes the Epipolar geometry relationship between two viewpoint images captured by general cameras.

We know that $p = K[I \quad 0]P$. By multiplying both sides by K^{-1}, we find that it can be transformed into the normalized camera: $K^{-1}p = K^{-1}K[I \quad 0]P = \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}P$.

Let us assume $p_c = K^{-1}p$, where p_c represents the corresponding point of P in the image under the normalized camera. Similarly, we have $p_c' = K'^{-1}p'$.

Based on the relationship defined by the essential matrix, we have:

$p_c'^{T}Ep_c = (K^{-1}p)^T T \times RK'^{-1}p' = p'^{T}K'^{-T}T \times RK^{-1}p = 0 \Rightarrow p'^{T}Fp = 0.$

Therefore, the fundamental matrix F is $K'^{-T}[T_\times]RK^{-1}$.

4.4 Estimate the Fundamental Matrix

Due to the existence of three degrees of freedom for rotation, three for translation, and one for the scale factor, the fundamental matrix FF has seven degrees of freedom. To solve FF, we select eight pairs of corresponding points. We write p, p', and F as: $p = \begin{bmatrix}u\\ v\\ 1\end{bmatrix}$, $p' = \begin{bmatrix}u'\\ v'\\ 1\end{bmatrix}$, $F = \begin{bmatrix}F_{11} & F_{12} & F_{13}\\ F_{21} & F_{22} & F_{23}\\ F_{31} & F_{32} & F_{33}\end{bmatrix}$. This yields:

$$(u', v', 1)\begin{pmatrix}F_{11} & F_{12} & F_{13}\\ F_{21} & F_{22} & F_{23}\\ F_{31} & F_{32} & F_{33}\end{pmatrix}\begin{pmatrix}u\\ v\\ 1\end{pmatrix} = 0 \Rightarrow (uu', vu', u', uv', vv', v', u, v, 1)\begin{pmatrix}F_{11}\\ F_{12}\\ F_{13}\\ F_{21}\\ F_{22}\\ F_{23}\\ F_{31}\\ F_{32}\\ F_{33}\end{pmatrix} = 0.$$

By combining the equations for the eight pairs of corresponding points, we can formulate a homogeneous linear system: $Wf = \begin{bmatrix}u_1u_1' & \cdots & 1\\ \vdots & \ddots & \vdots\\ u_8u_8' & \cdots & 1\end{bmatrix}\begin{bmatrix}F_{11}\\ \vdots\\ F_{33}\end{bmatrix} = 0.$

However, the least-squares solution for F, denoted as $\hat{F}$, typically has a rank of 3, whereas the rank of F must be 2. Therefore, we need to perform a singular value decomposition (SVD) on $\hat{F}$ and adjust it to enforce the rank-2 constraint, yielding the correct F.

The eight-point algorithm still suffers from a major issue: the numerical values of elements in W may vary significantly. To address this, we need to normalize the eight points. We apply a transformation T (a translation and scaling) to each image such that: The centroid of the points is moved to the origin; The mean squared distance of all points to the coordinate origin equals 2.

The complete normalized eight-point algorithm steps are as follows**:**

1. Compute the transformations T and T' for the two viewpoint images.
2. Normalize the eight pairs of corresponding points.
3. Use the eight-point algorithm to compute the fundamental matrix F_0.

4. Perform an inverse normalization on F_0 to obtain the final fundamental matrix F.

4.5 External Camera Parameters

With the correspondence between two views established, it becomes possible to solve for the currently unknown parameter R and T. As described earlier, $E = K_2^T F K_1$, and since K_1 and K_2 are obtained during camera calibration, the critical step is to decompose the essential matrix E.

First, we define two matrices, $W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $Z = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, where $Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} W = diag(1,1,0)W^T$. Rewrite $E = T \times R$ as $[T]_\times R$, where $[T]_\times$ can be expressed as $kUZU^T$, so if it don't consider the scale, $[T]_\times = Udiag(1,1,0)WU^T = Udiag(1,1,0)W^T U^T$.

Therefore:

$$E = [T]_\times R = Udiag(1,1,0)(WU^T R) \quad (1)$$

Perform a singular value decomposition (SVD) of E to obtain:

$$E = Udiag(1,1,0)V^T \quad (2)$$

By combining equations (1) and (2), we find:

$$R = UW^T V^T \text{and } R = UWV^T$$

Since it is necessary to ensure that the determinant of the rotation matrix is positive, so there have: $R = (detUWV^T)UWV^T \; or \; (detUWV^T)UW^T V^T$.

Since T satisfies $T \times T = 0$, we also have $[T]_\times T = 0$. After performing SVD, T can be expressed as:$T = \pm u_3$

As there are four possible combinations of R and T, during triangulation, the combination where the z-coordinates of the reconstructed points are positive in both camera coordinate systems should be chosen. This ensures the reconstructed points are in front of the cameras.

4.6 P3P Camera Pose

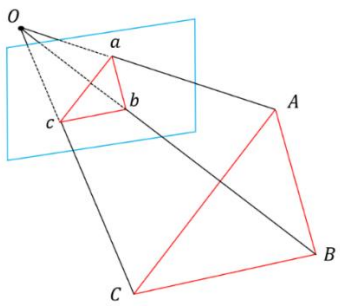

Figure 9 P3P

When the coordinates of points a, b, and c on the pixel plane and their corresponding 3D points A, B, and C in the world coordinate system are known, we first calculate the three angles using the directions of the lines $\overrightarrow{oa}$, $\overrightarrow{ob}$, and $\overrightarrow{oc}$. Next, we use the cosine law to find the lengths of OA, OB, and OC (since four possible solutions may be obtained, another corresponding point D is needed to represent the final OA, OB, and OC). By combining the lengths and directions of the three-line segments, we can determine the coordinates of points A, B, and C in the camera coordinate system. Finally, we compute the camera motion (R,T) based on the differences between the two coordinate systems.

4.7 Bundle Adjustment

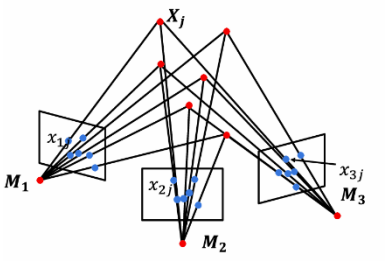

Figure 10 Bundle Adjustment

Since there is an error between the true value and the reconstruction point, the formula for the reprojection error is given: $E(M,X) = \sum_{i=1}^{m} \sum_{j=1}^{n} D(x_{ij}, M_i X_j)^2$.

The $E(M,X)$ minimization case is the best fit for the real value。

5 Data Pre-processing

In this section, we provide a detailed introduction to data collection and preprocessing.

5.1 Collect Data

Since the cheeks and mouth can obstruct the complete shape of the teeth, the first step after capturing the image is to use U-Net to segment the original image (Figure 11). This process results in an image that primarily contains the teeth.

Figure 11 Split teeth

5.2 Image Feature Points Extraction and Matching

5.2.1 Matching

Use SIFT to extract scale-invariant feature points from the images. After obtaining the feature points from all images, pair them up. For each feature point i in the left image, find the nearest feature point j_1 and the second nearest feature point j_2 in the right image, and record the distances d_1 and d_2 between j_1, j_2, and feature point i. Calculate the distance ratio ($\frac{d_1}{d_2}$); if this ratio is less than 0.6, then feature point i in the left image is considered a matching point with feature point j in the right image (Figure 12).

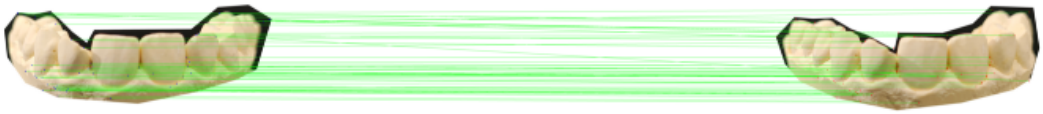

Figure 12 Matching point

5.2.2 RANSAC Estimate Fundamental Matrix and Find R, T

Input all matched point pairs obtained from the image matching process, and then randomly and uniformly sample pairs from these points. Based on the sampled eight-point pairs, use the eight-point algorithm to estimate the fundamental matrix F. Next, calculate whether the remaining points satisfy the current F (count the number of points that do satisfy it, which will be recorded as the score for the current F. Then, repeat the first three steps in sequence for a total of 100 times. The F with the highest score is the desired fundamental matrix. Then use F to obtain R and T.

6 Implementation Of 3D Reconstruction

6.1 Construction of Connected Graph

6.1.1 Tracks

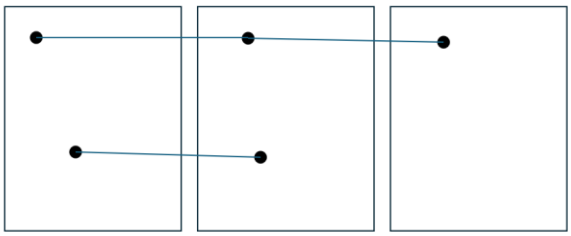

Figure 13 Tracks

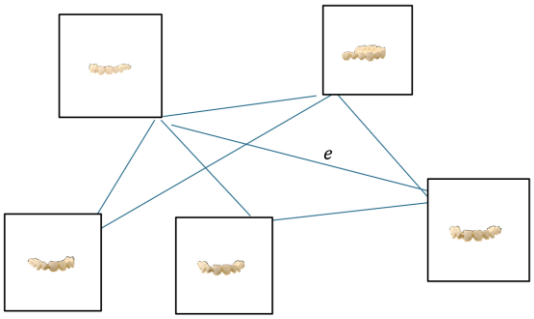

Figure 14 Graph

For a randomly selected point i, find the corresponding points in different images and connect them (Figure 13). The number of corresponding points is recorded as tracks (i.e., if point i is observed in three images, then tracks = 3). Tracks with a count of less than or equal to 2 will be discarded.

6.1.2 Graph

Construct a connected graph G (Figure 14), where an edge e is established between 2 images whenever the number of matching features between them exceeds 100.

6.2 Initial Point Cloud for Two-View Reconstruction

Select an edge e in G that satisfies the condition that the median of the angles between the rays when triangulating all corresponding points is greater than or equal to 3 degrees and less than or equal to 60 degrees. Next, robustly estimate the essential matrix E corresponding to e, and then determine the extrinsic parameters of the camera. Finally, triangulate the points that meet the condition $t \cap e$ to obtain the initial reconstruction result. Remove the edge e from G.

6.3 Multi-View Reconstruction

If there are still edges in G, continue selecting an edge e from G. The selection criterion is to choose the edge e that overlaps with the most already-reconstructed 3D points based on its tracks. Then, use the P3P method to estimate the camera's extrinsic parameters and triangulate new tracks. Finally, remove the edge e from G. Perform Bundle Adjustment at the end.

7 Result

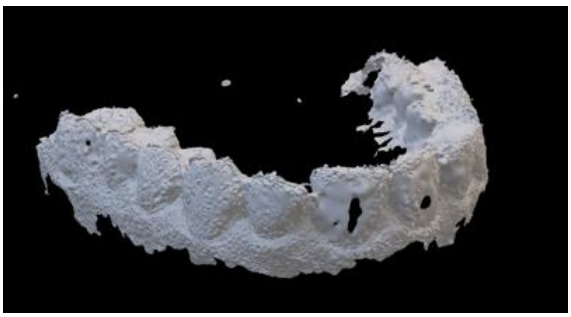

Figure 15 3D model

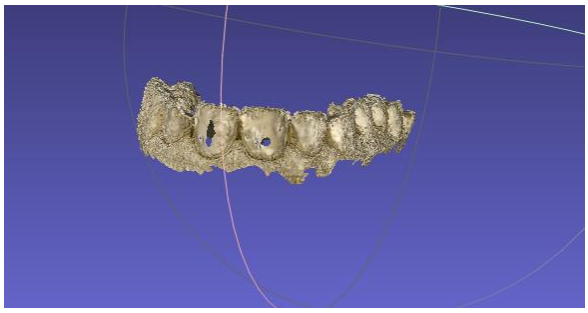

Figure 16 3D model

The teeth obtained after the Multi-View Reconstruction can be largely molded, well defined and almost complete in shape, although there are some imperfections, but they are sufficient for a cursory condition check. The disadvantages are the lack of precision, the inability to determine the color and the incompleteness of some of the teeth.

8 Discussion

The lack of dynamic adjustment of the algorithmic system to the process is a problem that can be solved. Moving the view more during the process of re-data collection to capture more 3D samples should improve the completeness of the reconstruction.

From the current results, it is difficult to obtain more accurate off-camera parameters for Multi-View Reconstruction due to the high overlap of teeth in the oral cavity and the small space. However, the transparency of the process and the ease of management are also points that make Multi-View Reconstruction superior to other methods.

Traditional methods use specialized medical equipment for imaging, which reduces spatial diversity. The use of mobile phones as a tool allows the collection of dental samples in increasingly decentralized scenarios, greatly increasing the flexibility of dental medical imaging. Leave this design as future work.

9 Related Work

9.1 Reconstruction of Teeth

In the field of dental 3D reconstruction, it is not practical to obtain complete multi-perspective images because of the limitation of clinical data collection, especially in the remote monitoring of orthodontic treatment. Current research attempts to reconstruct high-quality 3D dental models from sparse perspectives, such as 5 oral photographs. TeethDreamer [4] proposed a new framework based on large-scale diffusion model and neural surface reconstruction. For the sparse input perspective problem, the prior knowledge of diffusion model was used to generate multi-perspective images with known attitude, and the 3D perception feature attention mechanism is used to ensure the consistency of the generated perspectives. Subsequently, TeethDreamer combined with the loss of geometric sense normal significantly improved the accuracy of tooth shape and position reconstruction. Experimental results show that TeethDreamer is superior to the current mainstream methods in the task of tooth reconstruction.

Compared with the traditional method based on Structure-from-Motion (SFM) and multi-view stereo (MVS), TeethDreamer makes use of the advantage of depth learning model in sparse view reconstruction to solve the problem of missing view caused by insufficient input image. However, this method requires high quality and consistency of the generated image, and it relies on the training and reasoning process of the diffusion model, so the computational cost is relatively high.

9.2 Mobile phone imaging

In recent years, with the rapid development of mobile phone hardware and algorithms, mobile phone image has made great progress. The Efficient Hybrid Zoom using Camera Fusion on Mobile Phones [5] proposed by Wu et al. In this method, the image is fused by the machine learning model, and an adaptive hybrid algorithm is designed to deal with the depth mismatch, occlusion and error in the real scene. The system can process a 12-megapixel image in 500 milliseconds, making it clearer than ever before. With the development of research, mobile phone imaging will become better and better.

10 Conclusion

This paper presents a method of 3D dental reconstruction based on mobile phone sampling to solve the problems of visualization, dependence of fixed equipment and radiation risk in dental diagnosis. By using multi-view reconstruction technique, the oral image is segmented and processed accurately with convolutional neural network. The high-quality 3D dental model is reconstructed from the multi-tooth images taken by common mobile phone.

References

[1] L. de Agapito, E. Hayman, and I. Reid. Self-calibration of a rotating camera with varying intrinsic parameters. BMVC, 1998. 6.

[2] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N

[3] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. 2003.

[4] Xu, C., Liu, Z., Liu, Y., Dou, Y., Wu, J., Wang, J., Wang, M., Shen, D., & Cui, Z. (2024). TeethDreamer: 3D teeth reconstruction from five intra-oral photographs. *arXiv*. https://doi.org/10.48550/arXiv.2407.11419.

[5] Wu, X., Lai, W.-S., Shih, Y., Herrmann, C., Krainin, M., Sun, D., & Liang, C.-K. (2024). **Efficient hybrid zoom using camera fusion on mobile phones**. *arXiv*. https://doi.org/10.48550/arXiv.2401.01461.