ISSN 2959-6157

3D face recognition based on RGB-D data: a survey

Junhao Liu

Beijing University of Posts and Telecommunication, Beijing, China,100000

Abstract:

Face recognition, as a convenient, natural, and widely applied emerging technology, has achieved many significant research results in recent years. 2D face recognition has drawn extensive studies, while previously,2D face recognition is too sensitive to variations in features like facial expressions. To avoid the shortcoming, more attention was paid to the optimization of algorithms, stronger computational capabilities, and fusion strategies, which contributed greatly to the accuracy of face recognition and made it more outstanding. Compared to existing methods, RGB-D images tend to be more robust and reliable. Based on different processing methods of RGB-D 3D face data, researchers have proposed numerous 3D face recognition methods, such as 3D reconstruction methods from monocular RGB-D images, methods based on point cloud data, and methods based on image depth map data. This paper focuses mainly on the image depth map data method, analyzing its rich development history and its unique advantages and disadvantages in RGB-D 3D face recognition. Additionally, we introduced some common RGB-D face datasets, analyzing data collection methods.

Keywords: RGB-D face recognition, Feature-level fusion, Hybrid Fusion, Deep Learning Face Representation, 3D face data

1 Introduction

With the rapid development of artificial intelligence, face recognition technology has become one of the core technologies in many fields, such as identity verification, security monitoring, human-computer interaction, etc. Sukanya et al.[1]proposed a survey of several available recognition methods, revealing the very significance of human face recognition. Chihaou et al.[2]and Günther et al.[3] have given a relatively comprehensive summary of 2D face recognition. However, it's clear that traditional 2D face recognition methods have limitations in dealing with illumination changes, posture changes, and expression changes. In this regard, face recognition technology based on RGB-D data not only includes the texture information of color images (RGB) but also integrates the geometric shape information of depth images (D) by generating their corresponding depth maps, thus providing richer and more robust feature representations for face recognition.

3D morphable models(3DMM)[4] is a statistical model that obtains a low-dimensional representation of face shape and texture through principal component analysis of a large amount of 3D face data. Khan[5] compared diverse 3DMM techniques, which emphasize different issues, like texture estimation, shape-controlling limits, and so on. By combining RGB-D data with 3DMM, more accurate face shape estimation and texture recovery can be achieved.

Feature extraction methods, as one of the new methods for processing RGB-D data, have been proposed in recent years, such as Rahim et al.[6] attempted to supplement feature extraction with local binary patterns (LBPs), and some researchers extracted features from the geometrical shapes of the face in order to get sufficient detailed information, but in general, the accuracy of this method still needs to be improved.

The emergence of deep learning[100] has made face recognition methods more sensitive to changes in the face. It has also made various models smoother when dealing with conditions such as different noises, different lighting, and different occlusions. However, deep learning also brings many challenges to face recognition and faces in real conditions can be affected by low-quality data acquisition, posture deformation, and environmental changes. Wang et al. [95]provided a comprehensive review of the advanced progress in deep face recognition technology, covering a wide range of aspects, including algorithm design, protocol setting, and application scenarios. Liu et al.[96]classifies the data into two categories, virtual sample methods, and generic learning methods, based on different approaches to the data and analyses them separately. However, these are still not sufficiently comprehensive in their coverage. Zhou et al.[97]represented a detailed review of its history and categorized the frontier research into three different classes. Ning et al.[98]conducted a general overview of all sorts of face generation models, and the performance of existing models was examined

through experiments.

3D face reconstruction technology has been used in a wide range of fields, including plastic surgery and the entertainment industry, thanks largely to its advantageous features. These features may include a high degree of accuracy and realism. Cava et al.[94]reviewed existing 3D face reconstruction algorithms and analyzed both their obstacles and their achievements.Sharma et al.[99] Further review was made of various aspects of reconstruction methods, including deep learning, epipolar geometry, and so on.

In this paper, we introduce and analyze in detail the different ways of processing RGB-D data, focusing on face recognition based on 3DMM, face recognition based on feature extraction, face recognition based on deep learning, face recognition based on face reconstruction, as well as the commonly used datasets for training and validating the data, and give a summary and evaluation of the latest methods in each area, and make predictions and outlooks of the prospects, taking into account the development trends and the potential problems at present.

2 RGB-D 3D face

In recent years, with the continuous advancement of technology and the deep development of the computer vision field, the application of RGB-D information in 3D facial recognition has gained increasing attention. RGB images provide rich color and texture information.

2.1 3D Morphable FaceModel and Face Acquisition

In recent years, a variety of methods have emerged in the field of 3D facial modeling, dedicated to improving the robustness and performance of models. Zhong et al.[7] proposed an identity authentication method based on joint constraints, introducing central loss in model training to effectively capture individual expression, identity, posture, and lighting information, ensuring the authenticity of the generated faces. Luo et al.[8] used a random forest algorithm to estimate head models and integrate the optimal weights of facial vertices, thereby improving the robustness and accuracy of facial modeling. However, there are still deficiencies in reconstructing facial details. Tran et al.[9] used an approach of directly inputting unconstrained images, mapping them to 3D shapes and textures through a decoder in combination with projection parameters, achieving reconstruction of the original input faces. Compared to traditional 3DMM algorithms, this method is more compact and faithful, with a learning process that only requires weak supervision, offering more flexibility. On the other hand, Jiang et al.[10] noticed the limitation of 3DMM in visually discriminating face shapes after model reshaping and proposed the Shape Identity Recognition (SIR) metric, effectively addressing the lack of datasets simultaneously possessing identity information and 3D image data. Zhu et al.[11]created the FG3D database through a non-rigid ICP approach constrained by texture and combined it with the Fine-Grained Reconstruction Network (FGNET) to achieve fine-grained geometric reconstruction of faces, improving the precision of face acquisition. Li et al.[12], based on 3DMM, built a coarse-to-fine framework of deep neural networks, introducing semantic consistency constraints to improve the performance of 3D facial reconstruction and depth annotation. However, accurately depicting fine-grained facial details remains challenging. Lastly, Zhong et al.[13]established a 3D facial regression framework encoded by identity, expression, and posture parameters, employing a regression model learned from synthetic data based on CycleGAN. By applying triple constraints of joint embedding, depth imaging, and shape coherence in surface space, they effectively enhanced the coherence and robustness of 3D faces. These methods focus on improving the robustness and performance of 3D facial modeling, yet further progress is still needed for more significant improvements. While these methods concentrate on enhancing 3D facial modeling, further progress is deemed necessary for substantial improvements.

2.2 Feature Extraction and Classification

This technique extracts multidimensional features on RGB images, focusing on multimodal feature fusion and attention mechanisms. Dutta et al.[14] utilized a facial component mathematical model to generate basic facial components in four directions. They then extracted features from these components along with four other selected mixed components and applied genetic algorithms for the crossover fusion of feature vectors. This method not only improved the performance of target recognition but also significantly reduced the feature storage space required by traditional methods. Importantly, this method of feature selection could also be directly applied to result detection, ensuring accurate target identification. Zhu et al.[15] constructed attention fusion network, CMANet, which focuses on the role of soft masks and soft weights in the homomodal attention mechanism. Compared to traditional hard mask fusion methods, this approach better captures all potential information areas. Additionally, the network fuses multimodal features to ensure the unique advantages of each modality are fully utilized. Boumedine et al.[16] designed a 3D facial recognition system, which is based on extracting features in the normal direction using the SURF algorithm. By assigning the best weights to each component through the nearest classifier, they achieve feature fusion. Although this method performs excellently in terms of processing speed and accuracy, its adaptability to

occlusions and head tilts could be better, with a need for improved universality. Sui et al.[17] represented 3D scan images using three types of attribute maps and inputted the combination of attributes into the Efficient Feature Fusion System FFNET-M through multimodality. They also used two larger kernels, F3DNetsM and VGG16 masks, allowing FFNET-M to focus on both 2D and 3D spaces. This method excels in extracting 3D depth features and integrating 2D texture information, but the training cost is high. Wardeberg et al.[18]proposed a graph convolution-based feature extraction network for direct feature extraction on grids with a simple network model size. However, the mapping approach could be better than the method in terms of performance due to the training size problem for the inter-class separation problem. The method shows the potential of lattice-based feature extraction methods since it can have a more impressive recognition rate under datasets such as BU3DFE at 0.1% false acceptance rate. Sanyal et al's[19] RingNet(Fig.1) utilizes datasets of individuals that appear repeatedly and estimated 2D facial features. The loss method employed by the network provides the generator with discriminative capabilities, while the FLAME model separates expressions and other features in shape, making RingNet's parameter changes more flexible. This design gives RingNet high robustness under conditions of head illumination, occlusion, and more. Regarding these, feature extraction-based face recognition offers advantages such as effective feature capture and reduced computational complexity, but it may struggle with certain variations and incur high training costs.

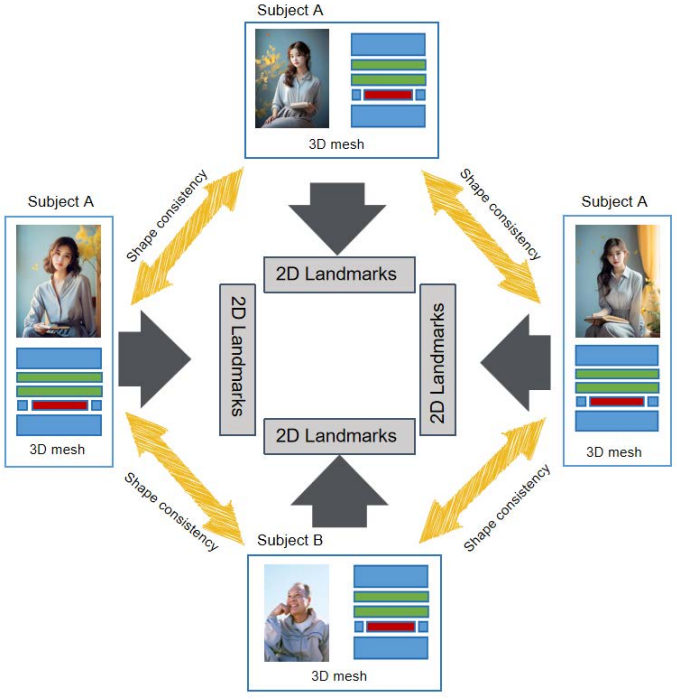

Fig.1: RingNet, taking multiple images of one single person(Subject A) and an image of a different person(Subject B) during training, maintains face consistency between the same object and shape inconsistency between various objects.

2.3 Deep Learning-Based Methods

With the flourishing development of deep learning technologies, the integration of RGB-D information with advanced algorithms such as neural networks has further propelled the performance enhancement of 3D facial recognition. Researchers have employed large-scale RGB-D datasets for model training and optimization, enabling the system to automatically learn more advanced and robust feature representations. This not only improves the accuracy of recognition but also allows the system to adapt more quickly to new scenes.

2.3.1 Based on Network Structure Optimization

This approach focuses on improving the performance and accuracy of facial recognition systems by continuously optimizing network structures and parameters. Zeng et al.[20] developed the Deep Fine-to-Fine Network (DF-2Net), which achieved significant progress. The network meticulously designed three modules and significantly optimized the effect of facial reconstruction through in-

dependent training, utilizing diverse data and training strategies. Especially in the task of converting a single facial image into a high-fidelity face, DF2Net demonstrated outstanding capabilities, achieving remarkable results. Meanwhile, Dutta et al.[21] introduced sparse principal component analysis during the convolution phase to learn multi-level filter banks, further combined with binary hash indexing and block histogram pooling techniques, and ultimately employed a linear support vector machine to classify the extracted features. The scale of this system can be flexibly adjusted according to the demand for accuracy. Notably, even with limited data, the system can still achieve high facial recognition accuracy. Additionally, its concise two-stage convolution design allows it to easily handle most scenarios. Li et al. [22]introduced normalization and hierarchical activation functions in CNNs, followed by the application of probabilistic max-pooling techniques to retain more feature representations. This method ensures accuracy but struggles with faces under adverse conditions. He et al.[23] incorporated fine-grained discriminators and wavelet-based discriminators into end-to-end deep networks, capable of generating high-resolution facial images and ensuring the integrity of facial images. He et al.[24]also attempted to introduce a self-supervised 3D facial reconstruction loss with two auxiliary functions in the 2D facial recognition pathway, forcing the FR to encode more facial depth information and reflectance information, which can serve as a baseline model for downstream tasks. Cui et al.[25] studied original strategies in fusion. They improved the performance of their method by weighing different fusion strategies and mixing them under specific rules. This method performed excellently on a self-built RGB-D dataset with stricter requirements for posture and lighting conditions, fully proving its effectiveness and practicality. Noting the huge potential in the ability and accuracy of point clouds to extract information, Atik et al.[26] uses the 3D features extracted from the point cloud to generate 2D images from which depth maps can be produced. This method eliminates the process of preprocessing the data and allows for scaling up the data, but the accuracy of face recognition is not satisfactory enough, and there is still much room for improvement. Hu et al.[27] collected a high-quality database, Extended-Multi-Dim, which includes color images, depth images, and 3D point clouds of each object. Through their standard protocol for fully utilizing features and image information, they demonstrated the feasibility of enhancing depth information quality to improve the accuracy of depth FR, thus improving the performance of models based on low-quality data through high-quality data. Lee et al.[28] constructed a comprehensive learning framework(Fig.2) capable of training multiple tasks simultaneously, including RGB facial parsing, depth facial parsing, and the mutual conversion between RGB and depth. This innovation achieved flexible conversion and end-to-end learning between RGB and depth parsing learning, providing strong support for supplementing annotated depth data. This method made significant progress in the diversity of facial parsing tasks and the utilization of depth data.

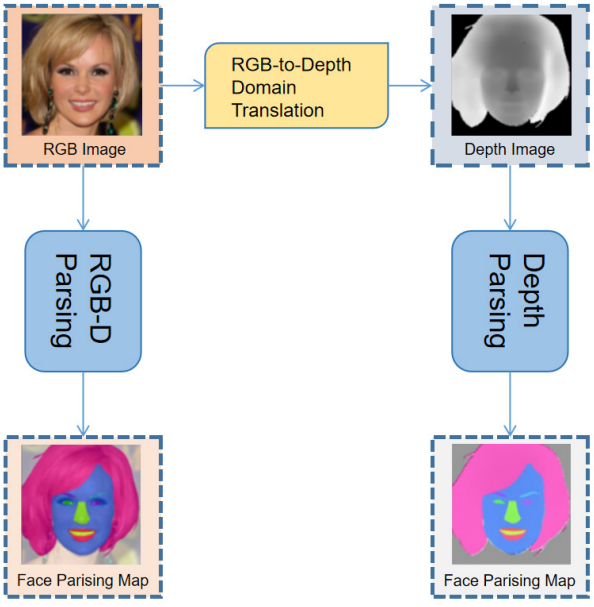

Fig.2: Learning Structure Lee Constructed and its results on face parsing, depth face parsing, and RGB-D Domain Translation

In the meantime, Li et al.[29] adopted a shape-based landmark detection method to align facial features. Through this alignment, they further trained transfer learning convolutional networks to adapt to diverse facial expressions. Ultimately, their convolutional neural network showed excellent performance in facial recognition tasks. This method not only reduced the data volume of the original 3D whole face but also made the convolutional network more efficient in network coding. Neto et al.[30] utilized a hybrid convolutional method based on shallow learning feature representation, suitable for both classification tasks and feature extraction modes. By combining manually created feature maps and 3D facial recognition using a constructed convolutional neural network, this method significantly reduced the model's training volume and computation time and displayed strong robustness against expression changes. This research provides a new solution for facial recognition under different expression changes. Niu et al.[31] proposed a 3D facial reconstruction method based on a single image. They reconstructed 3D faces from single images using the 3DMM method and obtained corresponding depth images. Combined with curvature feature extraction technology, this method efficiently extracts features from 3D faces and matches them with RGB images in the database to complete facial recognition. This method's outstanding advantage is its adaptability in low-light environments, though its accuracy in processing low-quality depth images still needs further improvement.

Deep networks exhibit significant superiority in feature extraction. Huang et al.[32]employed multiple networks for facial recognition, including feature restoration networks, feature extraction networks, and embedding matching modules, allowing for facial model preprocessing and improving the efficiency of facial recognition. Zhao et al.[33] proposed the Hierarchical Structure Lightweight Multi-Scale Fusion Network (LMFNet), combining mid and low-level adjacent layers with a Multi-Level Multi-Scale Feature Fusion (ML-MSFF) module, effectively extracting local information and combining it with global features obtained from the global convolutional network for a more comprehensive representation of faces. While this method improves the fluidity of human-machine interaction, it also reveals a vulnerability to noise interference. Khan et al.[34] contributed a deep neural estimation model, using a pre-trained network to initialize the encoder and simply inputting extracted features into a straightforward decoder to output high-quality facial depth maps. On public datasets, with high-quality training data, this method produces more accurate results. Compared to other depth estimation techniques, it shows significant improvements in volume and computational cost. Ding et al.[35] concatenated complementary facial features into high-order feature vectors and compressed dimensions through a three-layer stacked autoencoder (SAE), suitable for studying nonlinear dimensionality reduction. Hu et al. [36] constructed MCFLNet, which could enhance the performance of RGB-D FR under complex conditions by using a cross-modal learning constraint to extract specific modal features and modal shared features, utilizing a mutual restraint mechanism to extract complementary features. This method's effectiveness in handling multimodal data has been verified. Cai et al.[37] extracted complementary features under four overlapping facial component patches, refined specific facial structures with depth information and used a global descriptor to describe identity information. Additionally, the method compensates information using the upper half of the global descriptor and rigid local descriptor. This method has high recognition efficiency but is not sensitive enough to handle large posture changes.

To address the vulnerability of feature extraction methods to noise interference, Zeng et al. [38]normalized specific depth facial images and utilized a feature extraction network along with two convolutional neural networks to extract robust information, reducing the impact of deterministic factors such as noise, expressions, and lighting conditions. This method has a high generalization ability for the poses and expressions of recognized faces but has certain limitations in preserving facial identity information. Garg et al.[39] used a deep learning feature extraction and classification network, DeBNet, which filters all noise to make the system more stable. The network excels in recognizing photo impersonation, but due to the extensive features required, the training process is more time-consuming. In summary, these studies have made significant progress in the field of facial recognition and depth map generation, improving the accuracy, robustness, and efficiency through innovative methods and techniques. Gratis et al.[40] extracted RGBD patches around image points of interest, used CNNs, among others, to learn feature descriptors for facial patches, and applied the SRC algorithm to corresponding patches, finally generating the classification results through a score-level fusion scheme. This method achieved sufficient robustness on multiple benchmark RGBD databases. However, challenges and issues such as noise interference, pose changes and preservation of identity information still need to be addressed. Future research can focus on these areas to achieve further breakthroughs and progress.

In feature extraction, networks often incorporate attention mechanisms to improve performance. Uppal et al.[41] proposed an attention mechanism that effectively guides deep networks in extracting visual features from two modalities. By creating attention maps, this mechanism can focus on parts rich in features and salient parts, thereby

facilitating feature classification. In addition, this mechanism is versatile and can be combined with information from other features, guiding the network to focus on specific information. Zhang et al.[42] adopted a coarse supervision strategy to mitigate the common issues of noise and low resolution in depth maps. They incorporated a Branch Attention Module (BAM) and Edge Sensitive Attention Module (ESAM) into the refinement process for selective feature fusion and depth discontinuity restoration, respectively. Through this method, they were able to restore depth structures with finer details. However, the method often underperforms when depth maps are missing. Lin et al.[43] introduced an attention decomposition mechanism in the feature extraction process of mixed feature maps, which successfully decomposed features like pose and identity. They combined the Gradient Reversal Layer (GRL) with continuous index domain adaptation and constructed an embedded Convolutional Neural Network (eCNN) to simplify operations and reduce parameter size. This method effectively bridged the domain gap for frontal and profile faces, achieving a considerable recognition rate. Wang et al. [44]contributed a facial depth map transformer that progressively estimates facial depth maps based on RGB images and integrates facial depth maps through a Multi-Head Depth Attention (MDA) mechanism, thereby enhancing backbone features. This method facilitates facial image monitoring with facial depth maps, providing convenience for detecting low-quality compressed facial images.

Adversarial network models are commonly utilized in feature extraction. The adversarial algorithm MLAT proposed by Yu et al.[45] dynamically gives the corresponding adversarial samples according to the deep 3DFR model and uses a meta-learning framework to ensure the performance of the original algorithm to optimize the accuracy of 3DFR by obtaining a variety of adversarial samples in a round-robin optimization. This method can use fewer face data to optimize the model and get more considerable results. Jin et al.[46] designed a Generative Adversarial Network model (D+GAN) that generates higher quality facial depth maps, using NSST to transform and merge high-frequency and low-frequency sub-band images, then inverting the transformation to obtain the fused image, demonstrating strict robustness. Gecer et al.[47] designed an adversarial conversion method that does not require a large amount of paired data, based on a semi-supervised adversarial learning framework and ensemble-based loss, maintaining the pose, lighting, and other information of newly generated facial images, as well as preserving identity information, avoiding some flaws of adversarial conversion methods. Uppal et al.[48] proposed a Teacher-Student Adversarial Architecture (TS-GAN), significantly enhancing facial recognition performance through supervised mapping relationships between teacher and student components. However, this network might overlook other feature information when focusing on specific information, leading to a lack of integrity in the involved information. To address this issue, Uppal et al.[49] proposed a dual-layer attention mechanism for merging features from RGB and depth images. They sequentially used LSTM to learn the relationship between fused feature maps and utilized convolution to focus on spatial features. During the learning process, they used transfer learning from the training process of pure 2D RGB images as guidance, thereby achieving accurate results. However, this method incurs a high training cost.

Another drawback of feature extraction is the difficulty in handling blurry or flawed information. Ghosh et al.[50] proposed a representation learning algorithm that effectively establishes representational associations between RGB images and depth maps by integrating heterogeneous data and features. This algorithm showed a low dependence on depth maps in recognition tasks, producing feature-rich results that are particularly suitable for scenarios where depth map information is blurred or degraded. This research provides new insights into handling incomplete depth information. Zhu et al.[51] contributed a Progressive Multimodal Fusion Framework (PMMF) aimed at finely optimizing image details and rectifying errors in depth recognition for depth images captured by low-cost RGB-D cameras. By processing and recognizing depth maps and RGB images separately and then leveraging the implicit complementary information between them, PMMF combined the original fusion results, significantly enhancing the robustness of recognition information. This method has made significant progress in balancing the cost and accuracy of depth recognition. Lin et al.[52] proposed a two-stage pipeline(Fig.3) based on the pix2pix's deepened image capabilities and the Multi-Quality Fusion Network (MQFNet), optimized for high-quality depth images. By extracting and merging multi-level, multi-quality features from "conv" and "res" pipelines, this method excels in processing low-quality depth facial images. This research also provides a new solution for handling facial images of different qualities in depth.

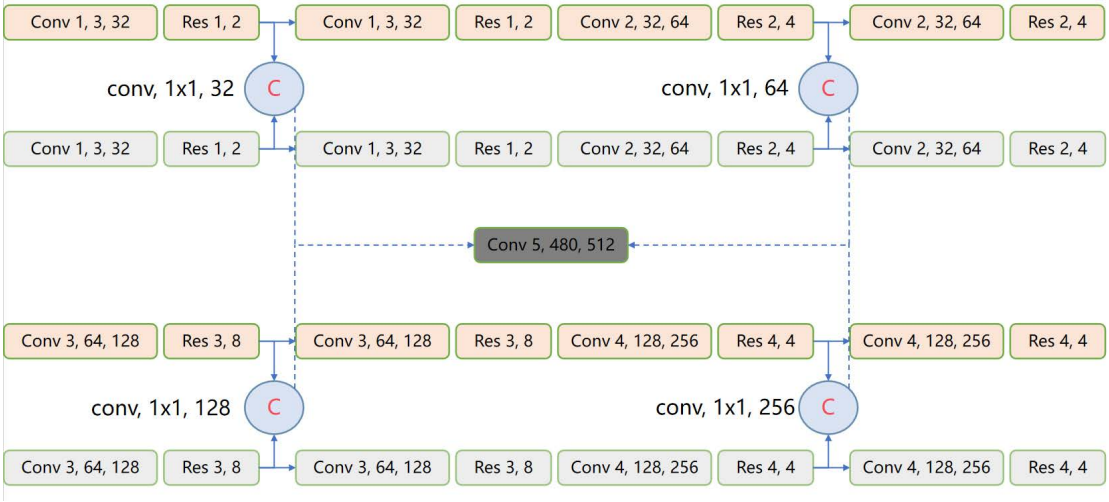

Fig.3: The two-stage pipeline architecture of MQFNet Lin proposed

In recent years, depth networks involved in RGB-D facial recognition have also touched on many other areas. Thamizharasan et al.[53] utilized conditional generative adversarial networks (cGANs) to extract gender and age from Structured Light (SL) and structure maps generated from SL, demonstrating the applicability of IRDP images in facial analysis. Pecoraro et al.[54] focused on the synergistic action of channel self-attention and convolution, proposing the Local Multi-Head Channel Self-Attention module (LHC: Local Multi-Head Channel self-attention). This channel-based module effectively utilizes the constant structure of images, applying convolution locally for facial recognition, providing new directions for the development of the computer vision field.

2.3.2 Based on Multi-Branch Structure

In the development of facial recognition, multi-branch processing, as a powerful mechanism, allows for analyzing facial data from multiple angles and levels. It enables the capture and fusion of information from different data sources, such as color images and depth images, achieving multi-level improvements in the accuracy of facial recognition. You et al.[55] calculated geometric information of 3D facial data based on discrete surface curvature flow and "flattened" it along with color information into a two-dimensional plane. By modifying the input layer of CNNs, they output images with nine channels. This multi-channel deep network showed considerable accuracy, but the model training process was complex. Lin et al.[56] designed dual CNN pathway networks for color images and depth images, respectively, using a designed loss function to strengthen two kinds of distinct, complementary features and fusing them beforehand to preserve these features. This method is beneficial for image classification. Zhao et al.[57] added a capsule network on top of a multi-scale local and global feature fusion module to strengthen 3D positional information, capturing more details in facial information and reducing redundant information in fused features. This method can provide more feature information for faces with missing features, but it still underperforms for low-quality 3D FR that changes over time. Zheng et al.[58] aimed to address the limitations of using low-resolution facial depth maps, proposing an edge-guided convolutional neural network that includes an edge prediction subnetwork and a depth reconstruction subnetwork. This method first predicts edges from the input for more accurate restoration, then uses concatenated features to reconstruct depth maps, accurately restoring high-frequency facial depth maps and revealing more edge details.

Neto et al.[59] proposed a flow attention network, effectively extracting feature information from deep images by incorporating self-attention and cross-attention mechanisms into the architecture, focusing the network's attention on facial information. This method facilitates better integration of all information from different data streams, including bridging the gap between high and low resolutions, making it easier to improve recognition rates in low-resolution features, and enhancing the robustness of recognition through different streams working collaboratively. Jiang et al.[60] utilized spatial and channel attention mechanisms, based on ResNet18, to create three branches that can capture RGB, depth map, and their fused modal features, performing a second step of feature extraction in a shared layer after a secondary fusion, applying complementary information contained in RGB and depth maps for end-to-end multimodal fusion facial recognition. This model performs well, but the overall process is cumbersome, and training costs are high. In summary, the facial recognition models supported by multi-branch structure deep networks excel in capturing details and

overall features. However, the current challenge lies in optimizing the training process of multi-branch networks to reduce computational costs while achieving comparable performance.

2.3.3 Based on Loss Function

In deep learning-based face recognition, well-designed loss functions can also effectively improve recognition accuracy and model generalization capabilities. Li et al.[61] replaced the three channels of RGB images with depth map channels, fusing RGB and depth maps before encoding and final recognition, i.e., MFEViT, introducing visual transformers in 2D+3D FER. This optimized structural features of RGB images and multimodal data significantly reduced the sample demand and mitigated the impact of noise on the training network, resulting in robust recognition results. Zheng et al.[62] concentrated on the limitations of existing multimodal fusion methods in recognizing different facial features, proposing a Complementary Multimodal Fusion Transformer (CMMF-Trans), which can supplement fused feature information. This transformer performs well on depth-encoded maps, especially under special conditions like shadows and lighting, and maintains good stability under extreme conditions. Khan et al.[63] proposed SOA neural architecture, which automatically collects the optimal parameters, reduces the model complexity in estimating the training model, enables the model to estimate the parameters for a single frame of a 2D image to estimate the local details, and normalizes the weights by introducing weights to estimate the depth residuals. Jiang et al.[64] regularized the distribution of features recognized with additional attributes, proposing an attribute-aware loss function based on CNN, making the training data collection more uniform and enhancing the model's accuracy. This method trains fewer models and has stronger versatility, promising significant research potential in the future. Zhang et al.[65] proposed conditional generative adversarial networks (cGANs) to fuse multimodal matching results in the process of reconstructing depth images, reducing the lossy compression of low-level information during data transmission to ensure result accuracy. In summary, by designing different loss functions, the accuracy and stability of facial recognition have been ensured, and the models' generalization capabilities have been enhanced.

2.4 3D Face Reconstruction

This method employs deep learning techniques to recover the three-dimensional shape and texture information of a face from input image or video data, which is then compared to produce recognition results. Kemelmacher et al.[66] focus on the global similarity of faces, extracting a universal shape from a single reference model and reconstructing the face by combining reflectance properties, illumination, depth values of boundary conditions, and multiple shadow information. This approach has low requirements for lighting conditions and can robustly generate pose information. Dib et al.[67]introduced a differentiable ray tracer into a CNN encoder for monocular face reconstruction. Under the method of deep network and differentiable network rendering ray tracing, high-quality lighting and BRDF are used to capture more detailed diffuse and specular reflections. This method can produce richer reflections, making the face reconstruction results more reliable under lighting conditions and the recognition results more robust.

To improve the accuracy of the reconstruction process, Xian et al.[68] Based on their method of using facial point clouds, facial landmarks are introduced in the reconstruction process to optimize details. They selectively smooth noise and holes in different areas of the face, making facial reconstruction more reliable. This method facilitates the direct acquisition of RGB-D images but has strict requirements for the depth range and is not conducive to processing multiple expressions. Petkova et al.[69] capture faces from multiple-view approaches, sequentially extracting and processing RGB and 3D features in parallel, and back-project facial key points. The 3D facial points obtained in this way are used for rough alignment steps, followed by a second processing to refine the results. This method achieves an average distance of less than 2 millimeters for 90% of the points between the generated model and the reference model when artificially generated data is used as input, demonstrating strong reconstruction capability. Hence, facial reconstruction that combines techniques like multiple shadow information, ray tracing, and facial landmarks can ensure the robustness of recognition results, but the method still poses certain challenges.

3 Experiments

3.1 RGB-D Facial Data

The AFW dataset[70], comprising 205 images housing 473 faces, meticulously annotates each face with a square bounding box, six key points, and three pose angles. Contrastingly, the AFLW dataset[71], a vast repository of 25,993 images spanning different color spaces, showcases a rich array of poses, expressions, and ethnicities, each annotated with 21-point annotations. The LFPW[72] dataset includes 1,432 images downloaded from the internet, containing 29-point annotations that enhance the positioning of the eyes and chin. Helen dataset[73] contains 2,330 images with highly detailed, consistent, and accurate main facial components, annotated with 68 points, which is the

most commonly used annotation scheme today.

The LFW[74] database, a formidable collection of 13,233 facial images representing 5,749 individuals, captures a broad spectrum of poses, lighting conditions, expressions, and age variations, providing a comprehensive resource for facial recognition research. The CASIA-WebFace[75] encompasses 494,414 facial images of 10,575 individuals, including diverse angles, nationalities, and lighting information. The Celebrities in Frontal-Profile in the Wild (CFP)[76]is a smaller dataset that collects 7,000 images of 500 people, including both frontal and profile images with random expressions. The Public Figures Face Database (PubFig)[77] is a dataset collected from internet search engines, featuring 58,797 images of 200 celebrities with different facial features and lighting conditions, providing corresponding annotation information and sampling strategies for ease of analysis. The CAS-PEAL database[78] comprises 99,594 photos of 1,040 Chinese people, showcasing variations in orientation, expression, accessories, etc., with 21 different head postures. WebFace260M dateset[79], as the largest facial database to date, consists of 260 million faces from 4 million identities, containing a large amount of noisy data. The high-quality data obtained through automatic cleaning resulted in the WebFace42M dataset, comprising 42 million images from 2.06 million identities, making it the largest public facial recognition dataset currently available.

WiderFace dateset[80] includes 32,203 images and 393,703 annotated faces, with training, validation, and test sets that span a wide range of scales, poses, lighting conditions, expressions, and occlusions. The IJB-A dataset[81] includes images of 500 subjects, totaling 5,712 images, with a wide range of facial poses, image resolutions, and lighting conditions, introducing set-to-set matching composed of heterogeneous media, with subjects from around the world showing rich diversity. CelebA dataset[82], with 202,599 facial images of 10,177 celebrities, each annotated with five landmark coordinates, includes 40 different attribute annotations such as glasses, black hair, etc. LS3DFace's data, gathered from multiple public datasets, includes 31,860 corresponding images for 3,153 subjects. UoY datebase[83] collected over 5,000 sets of color images and 3D mesh data using a high-density structured light camera, including information on more than 300 subjects. The Bosphorus dataset[84] encompasses images of 105 subjects with up to 4,666 images, faces equipped with up to 34 expressions, and up to 13 postures. UMBDB[85] contains many faces obstructed by objects like hats and scarves, with 1,473 pairs of depth and color images for 143 subjects used for testing 3D facial analysis with obstructions. 3DTEC dataset[86] consists of 428 images from 241 subjects, including 107 pairs of highly similar twin faces. Texas-3D dataset[87] includes 1,149 images of 118 people, each annotated with their gender, ethnicity, facial expressions, etc., and several artificially located facial anthropometric coordinate points.FRGC-v2.0 dataset's[88] 4,007 images of 466 subjects in 3D verification settings offer insights into the challenges and opportunities posed by three-dimensional facial recognition. Most of the 5,711 images of 509 individuals contained in the Lock3DFace dateset[89] contain relatively strong noise, with a wide span of variation in pose, demeanor, occlusion, etc. F3D-FD[90] collected multi-angle images of 2,476 individuals, which were not impeded. IIIT-D[91] collected 4,605 images of 106 subjects using Kinect, which contained a wide variety of demeanor variations.KinectFaceDB[92]collection of faces captured at different periods with diverse occlusions and lighting conditions offers insights into the robustness of facial recognition algorithms in dynamic environments. BU-3DFE[93] contains 2,500 images with various facial expressions collected from 100 subjects.

Table 1:Common RGB-D databases

Name of dataset	occlusion	expression	Number of images	Number of identity
AFW	No	No	205	479
AFLW	No	No	25993	-
LFPW	No	No	1432	-
Helen	No	No	2330	-
LFW	Yes	Yes	13233	5749
CASIA-WebFace	No	No	494414	10575
CFP	No	Yes	7000	500
Pubg	No	Yes	58797	200
CAS-PEAL	No	Yes	99594	1040

Name of dataset	occlusion	expression	Number of images	Number of identity
WebFace260M	Yes	No	260M	4M
WebFace42M	Yes	No	42M	2M
Wider Face	Yes	Yes	32203	393703
IJB-A	No	No	5712	500
CelebA	Yes	Yes	202599	10177
LS3DFace	No	No	31860	3153
UoY	No	Yes	5000	350
Bosphorus	Yes	Yes	4666	105
UMBDB	Yes	Yes	1473	143
3DTEC	No	Yes	428	241
Texas-3D	No	Yes	1149	118
FRGC-v2.0	No	Yes	4950	466
Lock3DFace	Yes	Yes	5711	509
F3D-FD	Yes	No	-	2476
IIIT-D	Yes	Yes	4605	106
KinectFaceDB	Yes	Yes	936	52
BU-3DFE	No	Yes	2500	100

Data Scale, Collection Methods,

3.2 Evaluation Metrics

There are various evaluation standards for facial recognition accuracy. Accuracy (Acc) is often used as an evaluation metric for facial recognition. It is defined as:

Data Scale, Collection Methods,

3.2 Evaluation Metrics

There are various evaluation standards for facial recognition accuracy. Accuracy (Acc) is often used as an evaluation metric for facial recognition. It is defined as:

$$Acc = \frac{TA + TR}{TA + TR + FA + FR} \tag{1}$$

Where TA and TR represent the true accept and true reject instances, respectively, while FA and FR represent false accept and false reject instances. Its specific meaning is the proportion of correct facial pairs out of all tested facial pairs. However, due to the extreme imbalance of facial pairs in practice, Acc is rarely directly used to evaluate the accuracy of facial recognition models.

The F1 score is the harmonic mean of precision (P) and recall (R), which can also effectively assess the accuracy of facial recognition models. The definitions of precision and recall are as follows:

$$P = \frac{TP}{TP + FP} \tag{2}$$

(3)

Where TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively. Moreover, the ROC curve is often used to evaluate the performance of algorithms. The higher the slope of the ROC curve, the higher the classification accuracy the model can achieve in case of misjudgment. The Area Under the Curve (AUC) represents a comprehensive indicator of classifier performance; generally, the larger the AUC value, the better the classifier's performance. Because AUC demonstrates the robustness of classifier performance at different thresholds, it is widely used in the evaluation of classification models in computer vision. The ROC is constructed as follows: We define the False Positive Rate (FPR) and True Positive Rate (TPR) as follows:

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

Where TA and TR represent the true accept and true reject instances, respectively, while FA and FR represent false accept and false reject instances. Its specific meaning is the proportion of correct facial pairs out of all tested facial pairs. However, due to the extreme imbalance of facial pairs in practice, Acc is rarely directly used to evaluate the accuracy of facial recognition models.

Based on the algorithm's output results and the true labels,

a set of FPR and TPR values under different classification thresholds can be obtained. Plotting FPR on the x-axis and TPR on the y-axis through a series of points approximates the ROC curve.

3.3 Results

Following are the methods that are commonly used with accuracy respectively:

Table 2:Results tested in common databases

Methods	Database	ACC	Time of publication	Regarding
MFEVIT	BU-FE3D	90.28	2021	[84]
	Bosphorus	89.72		
CCs	Bosphorus	98.68	2020	[67]
	Texas3D	99.19		
FFNet	BU-FE3D	89.82	2021	[69]
	Bosphorus	87.65		
MCFLNet	Lock3DFace	97.78	2023	[21]
	IIIT-D	99.89		
Uppal et al.	CurtinFaces	99.1	2020	[23]
	Lock3DFace	87.3		
	IIIT-D	99.7		
DSNet	KinectFaceDB	96.3	2023	[26]
	IIIT-D	81.29		
Wardeberg et al.	FRGCv2	99.55	2021	[45]
	BU-FE3D	87.6		
	Bosphorus	86.0		
AFNet-M	BU-FE3D	90.08	2023	[49]
	Bosphorus	88.31		
MLAT-PointNet+	FRGCv2	98.01	2022	[47]
	Bosphorus	99.78		
	BU-FE3D	100		
Uppal et al.	Lock3DFace	87.3	2021	[73]
	CurtinFaces	99.1		
	IIIT-D	99.7		
	KaspAROV	95.3		
SpPCANet-1	Frav3D	96.93	2020	[71]
	Bosphorus	98.54		
	Casia3D	88.75		
Chiu et al.	BU-FE3D	100	2023	[74]
	Texas3D	100		
	Bosphorus	100		
LMFNet	KinectFaceDB	94.90	2023	[82]
	Bosphorus	93.03		
	Lock3DFace	88.01		
SMM	IIIT-D	99.61	2021	[81]

Methods	Database	ACC	Time of publication	Regarding
You et al.	Bosphorus	98.6	2020	[34]
	Texas	98.6		
Neto et al.	Bosphorus	93.54	2020	[35]
	KinectFaceDB	96.1		
Niu et al.	Lock3DFace	78.24	2023	[44]
	FRGCv2	83.39		
	Texas3D	84.25		
PMMF	Lock3DFace	90.9	2023	[76]
	IIIT-D	99.8		

4. Conclusion and future work

This paper has summarized and analyzed various advances in the current advanced face recognition methods based on RGB-D data in a more comprehensive way, compared the performance of many of the important methods involved on different datasets, and identified problems in the results, which can be used to point out the direction for the future development of face recognition.

Predictably, the current problems facing 3D RGB-D face recognition are as follows:

- 1. Low-quality RGB-D data

At present, the collection of RGB-D data is greatly affected by various factors such as local hardware limitations, resolution, and noise level, so it is considered difficult to achieve large-scale, high-quality quantity collection, and there are fewer large-scale face databases in the public domain, which drastically limits the universality of model training. Different datasets handle data in different ways, making it more difficult to generalize across datasets.

- 2. Interference from variable factors

In real-world conditions, various variables such as occlusion, lighting, posture, and expression tend to bias face data. Although there are a large number of algorithms that deal with these variables in different ways to minimize their impact, a certain amount of information loss is unavoidable, and the size of the model algorithms also increases, which is a major challenge for further improving the robustness and efficiency of face recognition.

- 3.Difficulty in Multimodal fusion

In deep learning-based data processing, the depth map with the RGB map are often fused in different modalities, which can provide a lot of information that is difficult to obtain in a single modality and can also extract more detailed and specific features; however, it is inevitable to produce redundant information between different modalities, and the algorithms can easily become more complex. In this case, a good and feasible fusion strategy can be designed to allow the computer to extract facial features quickly without wasting computational resources.

Overall, face recognition has made considerable and even remarkable achievements in many aspects. However, it still needs to be tested and explored in more aspects, and we expect that face recognition will overcome the current problems and become an indispensable part of society in the future.

References

[1]Sukanya C M, Gokul R, Paul V. A survey on object recognition methods[J]. International Journal of Science, Engineering and Computer Technology, 2016, 6(1): 48.

[2]Chihaoui M, Elkefi A, Bellil W, et al. A survey of 2D face recognition techniques[J]. Computers, 2016, 5(4): 21.

[3]Günther M, El Shafey L, Marcel S. 2D face recognition: An experimental and reproducible research survey[J]. 2017.

[4]Blanz V, Vetter T. Face recognition based on fitting a 3D morphable model[J]. IEEE Transactions on pattern analysis and machine intelligence, 2003, 25(9): 1063-1074.

[5] Khan M S, Jehanzeb M, Babar M I, et al. Face Recognition Analysis Using 3D Model[C]//Emerging Technologies in Computing: First International Conference, iCETiC 2018, London, UK, August 23–24, 2018, Proceedings 1. Springer International Publishing, 2018: 220-236.

[6]Abiantun R, Prabhu U, Savvides M. Sparse feature extraction for pose-tolerant face recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(10): 2061-2073.

[7]Zhong Y, Pei Y, Li P, et al. Face denoising and 3D reconstruction from a single depth image[C]//2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020). IEEE, 2020: 117-124.

[8]Luo C, Zhang J, Bao C, et al. Robust 3D face modeling and tracking from RGB-D images[J]. Multimedia Systems, 2022, 28(5): 1657-1666.

[9]Tran L, Liu X. Nonlinear 3d face morphable model[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7346-7355.

[10]Jiang D, Jin Y, Zhang F L, et al. Reconstructing recognizable

3d face shapes based on 3d morphable models[C]//Computer Graphics Forum. 2022, 41(6): 348-364.
[11]Zhu X, Yang F, Huang D, et al. Beyond 3dmm space: Towards fine-grained 3d face reconstruction[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. Springer International Publishing, 2020: 343-358.
[12]Li P, Pei Y, Zhong Y, et al. Robust 3D face reconstruction from single noisy depth image through semantic consistency[J]. IET Computer Vision, 2021, 15(6): 393-404.
[13]Zhong Y, Pei Y, Li P, et al. Face denoising and 3D reconstruction from a single depth image[C]//2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020). IEEE, 2020: 117-124.
[14]Dutta K, Bhattacharjee D, Nasipuri M, et al. Complement component face space for 3D face recognition from range images[J]. Applied Intelligence, 2021, 51(4): 2500-2517.
[15]Zhu Z, Sui M, Li H, et al. CMANET: Curvature-Aware Soft Mask Guided Attention Fusion Network for 2D+ 3D Facial Expression Recognition[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022: 1-6.
[16]Boumedine A Y, Bentaieb S, OUAMRI A. 3D Face Identification based on Normal Maps[C]//Proceedings of the International Conference on Advances in Communication Technology, Computing and Engineering, Meknes. 2022: 260-269.
[17]Sui M, Zhu Z, Zhao F, et al. FFNet-M: Feature fusion network with masks for multimodal facial expression recognition[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021: 1-6.
[18]Wardeberg H. Mesh-based 3D face recognition using Geometric Deep learning[D]. NTNU, 2021.
[19]Sanyal S, Bolkart T, Feng H, et al. Learning to regress 3D face shape and expression from an image without 3D supervision[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7763-7772.
[20]Zeng X, Peng X, Qiao Y. Df2net: A dense-fine-finer network for detailed 3d face reconstruction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2315-2324.
[21]Dutta K, Bhattacharjee D, Nasipuri M. SpPCANet: A simple deep learning-based feature extraction approach for 3D face recognition[J]. Multimedia Tools and Applications, 2020, 79(41): 31329-31352.
[22]Li J, Qiu T, Wen C, et al. Robust face recognition using the deep C2D-CNN model based on decision-level fusion[J]. Sensors, 2018, 18(7): 2080.
[23]He R, Cao J, Song L, et al. Adversarial cross-spectral face completion for NIR-VIS face recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(5): 1025-1037.
[24]He M, Zhang J, Shan S, et al. Enhancing face recognition with self-supervised 3d reconstruction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4062-4071.
[25]Cui J, Han H, Shan S, et al. RGB-D face recognition: A comparative study of representative fusion schemes[C]//Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13. Springer International Publishing, 2018: 358-366.
[26]Atik M E, Duran Z. Deep learning-based 3D face recognition using derived features from point cloud[C]//Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications. Springer International Publishing, 2021: 797-808.
[27]Hu Z, Gui P, Feng Z, et al. Boosting depth-based face recognition from a quality perspective[J]. Sensors, 2019, 19(19): 4124.
[28]Lee J, Bhattarai B, Kim T K. Face parsing from RGB and depth using cross-domain mutual learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1501-1510.
[29]Li Z, Zou H, Sun X, et al. 3d expression-invariant face verification based on transfer learning and siamese network for small sample size[J]. Electronics, 2021, 10(17): 2128.
[30]Cardia Neto J B. 3D face recognition with descriptor images and shallow convolutional neural networks[J]. 2020.
[31]Feng Z, Zhao Q. Robust face recognition with deeply normalized depth images[C]//Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13. Springer International Publishing, 2018: 418-427.
[32]Huang Y H, Chen H H. Deep face recognition for dim images[J]. Pattern Recognition, 2022, 126: 108580.
[33]Zhao P, Ming Y, Meng X, et al. LMFNet: A lightweight multiscale fusion network with hierarchical structure for low-quality 3-D face recognition[J]. IEEE Transactions on Human-Machine Systems, 2022, 53(1): 239-252.
[34]Khan F, Shariff W, Farooq M A, et al. A robust lightweight fused-feature encoder-decoder model for monocular facial depth estimation from single images trained on synthetic data[J]. IEEE Access, 2023.
[35]Ding C, Tao D. Robust face recognition via multimodal deep face representation[J]. IEEE transactions on Multimedia, 2015, 17(11): 2049-2058.
[36]Hu W. Improving 2D face recognition via fine-level facial depth generation and RGB-D complementary feature learning[J]. arXiv preprint arXiv:2305.04426, 2023.
[37]Cai Y, Lei Y, Yang M, et al. A fast and robust 3D face recognition approach based on deeply learned face representation[J]. Neurocomputing, 2019, 363: 375-397.
[38]Zeng X, Peng X, Qiao Y. Df2net: A dense-fine-finer network for detailed 3d face reconstruction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2315-2324.
[39]Garg S, Mittal S, Kumar P, et al. DeBNet: multilayer deep network for liveness detection in face recognition system[C]//2020 7th International Conference on Signal

Processing and Integrated Networks (SPIN). IEEE, 2020: 1136-1141.
[40]Grati N, Ben-Hamadou A, Hammami M. Learning local representations for scalable RGB-D face recognition[J]. Expert Systems With Applications, 2020, 150: 113319.
[41]Uppal H, Sepas-Moghaddam A, Greenspan M, et al. Depth as attention for face representation learning[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 2461-2476.
[42]Zhang F, Liu N, Duan F. Coarse-to-Fine depth super-resolution with adaptive RGB-D feature attention[J]. IEEE Transactions on Multimedia, 2023.
[43]Lin W C, Chiu C T, Shih K C. RGB-D Based Pose-Invariant Face Recognition Via Attention Decomposition Module[C]// ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.
[44]Wang H, Li M, Li S, et al. Exploring Depth Information for Face Manipulation Detection[J]. arXiv preprint arXiv:2212.14230, 2022.
[45]Yu C, Zhang Z, Li H, et al. Meta-learning-based adversarial training for deep 3D face recognition on point clouds[J]. Pattern Recognition, 2023, 134: 109065.
[46]Jin B, Cruz L, Goncalves N. Pseudo RGB-D face recognition[J]. IEEE Sensors Journal, 2022, 22(22): 21780-21794.
[47]Gecer B, Bhattarai B, Kittler J, et al. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 217-234.
[48]Uppal H. Attention and Depth Hallucination for RGB-D Face Recognition with Deep Learning[D]. Queen's University (Canada), 2021.
[49]Uppal H, Sepas-Moghaddam A, Greenspan M, et al. Two-level attention-based fusion learning for RGB-D face recognition[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 10120-10127.
[50]Ghosh S, Singh R, Vatsa M, et al. RGB-D face recognition using reconstruction based shared representation[C]//2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). IEEE, 2021: 1-8.
[51]Zhu Y, Gao J, Wu T, et al. Exploiting enhanced and robust RGB-D face representation via progressive multimodal learning[J]. Pattern Recognition Letters, 2023, 166: 38-45.
[52]Lin S, Jiang C, Liu F, et al. High quality facial data synthesis and fusion for 3D low-quality face recognition[C]//2021 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2021: 1-8.
[53]Thamizharasan V, Das A, Battaglino D, et al. Face attribute analysis from s
[54]Pecoraro R, Basile V, Bono V. Local multi-head channel self-attention for facial expression recognition[J]. Information, 2022, 13(9): 419. tructured light: an end-to-end approach[J]. Multimedia Tools and Applications, 2023, 82(7): 10471-10490.
[55]You Z, Yang T, Jin M. Multi-channel deep 3D face recognition[J]. arXiv preprint arXiv:2009.14743, 2020.
[56]Lin T Y, Chiu C T, Tang C T. RGB-D based multimodal deep learning for face identification[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 1668-1672.
[57]Zhao P, Ming Y, Hu N, et al. DSNet: Dual-stream multi-scale fusion network for low-quality 3D face recognition[J]. AIP Advances, 2023, 13(8).
[58]Zhang F, Liu N, Chang L, et al. Edge-guided single facial depth map super-resolution using CNN[J]. IET Image Processing, 2020, 14(17): 4708-4716.
[59]Cardia Neto J B. 3D face recognition with descriptor images and shallow convolutional neural networks[J]. 2020.
[60]Jiang L, Zhang J, Li C, et al. Rgb-d face recognition via spatial and channel attentions[C]//2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2021, 5: 2037-2041.
[61]Li H, Sui M, Zhu Z, et al. MFEViT: A Robust Lightweight Transformer-based Network for Multimodal 2D+ 3D Facial Expression Recognition[J]. arXiv preprint arXiv:2109.13086, 2021.
[62]Zheng H, Wang W, Wen F, et al. A complementary fusion strategy for rgb-d face recognition[C]//International Conference on Multimedia Modeling. Cham: Springer International Publishing, 2022: 339-351.
[63]Khan F, Farooq M A, Shariff W, et al. Towards monocular neural facial depth estimation: Past, present, and future[J]. IEEE Access, 2022, 10: 29589-29611.
[64]Jiang L, Zhang J, Deng B. Robust RGB-D face recognition using attribute-aware loss[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(10): 2552-2566.
[65]Zhang W, Shu Z, Samaras D, et al. Improving heterogeneous face recognition with conditional adversarial networks[J]. arXiv preprint arXiv:1709.02848, 2017.
[66]Kemelmacher-Shlizerman I, Basri R. 3D face reconstruction from a single image using a single reference face shape[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 33(2): 394-405.
[67]Dib A, Thebault C, Ahn J, et al. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 12819-12829.
[68]Xiao M, Yi H, Huang Y, et al. Effective Key Region-Guided Face Detail Optimization Algorithm for 3D Face Reconstruction[J]. Journal of Sensors, 2022, 2022.
[69]Petkova R, Manolova A, Tonchev K, et al. 3D face reconstruction and verification using multi-view RGB-D data[C]//2022 Global Conference on Wireless and Optical Technologies (GCWOT). IEEE, 2022: 1-6.
[70]Koestinger M, Wohlhart P, Roth P M, et al. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization[C]//2011 IEEE international

conference on computer vision workshops (ICCV workshops). IEEE, 2011: 2144-2151.
[71]Devi P R S, Baskaran R. SL2E-AFRE: Personalized 3D face reconstruction using autoencoder with simultaneous subspace learning and landmark estimation[J]. Applied Intelligence, 2021, 51(4): 2253-2268.
[72]Belhumeur P N, Jacobs D W, Kriegman D J, et al. Localizing parts of faces using a consensus of exemplars[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(12): 2930-2940.
[73]Le V, Brandt J, Lin Z, et al. Interactive facial feature localization[C]//Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12. Springer Berlin Heidelberg, 2012: 679-692.
[74]Huang G B, Learned-Miller E. Labeled faces in the wild: Updates and new reporting procedures[J]. Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep, 2014, 14(003).
[75]Yi D, Lei Z, Liao S, et al. Learning face representation from scratch[J]. arXiv preprint arXiv:1411.7923, 2014.
[76]Sengupta S, Chen J C, Castillo C, et al. Frontal to profile face verification in the wild[C]//2016 IEEE winter conference on applications of computer vision (WACV). IEEE, 2016: 1-9.
[77]Kumar N, Berg A C, Belhumeur P N, et al. Attribute and simile classifiers for face verification[C]//2009 IEEE 12th international conference on computer vision. IEEE, 2009: 365-372.
[78]Gao W, Cao B, Shan S, et al. The CAS-PEAL large-scale Chinese face database and baseline evaluations[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2007, 38(1): 149-161.
[79]Zhu Z, Huang G, Deng J, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10492-10502.
[80]Yang S, Luo P, Loy C C, et al. Wider face: A face detection benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5525-5533.
[81]Klare B F, Klein B, Taborsky E, et al. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1931-1939.
[82]Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3730-3738
[83]Heseltine T, Pears N, Austin J. Three-dimensional face recognition using combinations of surface feature map subspace components[J]. Image and Vision Computing, 2008, 26(3): 382-396.
[84]Savran A, Alyüz N, Dibeklioğlu H, et al. Bosphorus database for 3D face analysis[C]//Biometrics and Identity Management: First European Workshop, BIOID 2008, Roskilde, Denmark, May 7-9, 2008. Revised Selected Papers 1. Springer Berlin Heidelberg, 2008: 47-56.
[85]Colombo A, Cusano C, Schettini R. UMB-DB: A database of partially occluded 3D faces[C]//2011 IEEE international conference on computer vision workshops (ICCV workshops). IEEE, 2011: 2113-2119.
[86]Vijayan V, Bowyer K W, Flynn P J, et al. Twins 3D face recognition challenge[C]//2011 international joint conference on biometrics (IJCB). IEEE, 2011: 1-7.
[87]Gupta S, Castleman K R, Markey M K, et al. Texas 3D face recognition database[C]//2010 IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI). IEEE, 2010: 97-100.
[88]Phillips P J, Flynn P J, Scruggs T, et al. Overview of the face recognition grand challenge[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 2005, 1: 947-954.
[89]Zhang J, Huang D, Wang Y, et al. Lock3dface: A large-scale database of low-cost kinect 3d faces[C]//2016 International Conference on Biometrics (ICB). IEEE, 2016: 1-8.
[90]Urbanová P, Ferková Z, Jandová M, et al. Introducing the FIDENTIS 3D face database[J]. AnthropologicAl review, 2018, 81(2): 202-223.
[91]Goswami G, Bharadwaj S, Vatsa M, et al. On RGB-D face recognition using Kinect[C]//2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, 2013: 1-6.
[92]Min R, Kose N, Dugelay J L. Kinectfacedb: A kinect database for face recognition[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2014, 44(11): 1534-1548.
[93]Yin L, Wei X, Sun Y, et al. A 3D facial expression database for facial behavior research[C]//7th international conference on automatic face and gesture recognition (FGR06). IEEE, 2006: 211-216.
[94]La Cava S M, Orrù G, Goldmann T, et al. 3D face reconstruction for forensic recognition-a survey[C]//2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022: 930-937.
[95]Wang M, Deng W. Deep face recognition: A survey[J]. Neurocomputing, 2021, 429: 215-244.
[96]Liu F, Chen D, Wang F, et al. Deep learning based single sample face recognition: a survey[J]. Artificial Intelligence Review, 2023, 56(3): 2723-2748.
[97]Zhou S, Xiao S. 3D face recognition: a survey[J]. Human-centric Computing and Information Sciences, 2018, 8(1): 35.
[98]Ning X, Nan F, Xu S, et al. Multi-view frontal face image generation: a survey[J]. Concurrency and Computation: Practice and Experience, 2023, 35(18): e6147.
[99]Sharma S, Kumar V. 3d face reconstruction in deep learning era: A survey[J]. Archives of Computational Methods in Engineering, 2022, 29(5): 3475-3507.
[100]Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.