Patterns

A fine-grained network for human identification using panoramic dental images

Highlights

- Our key idea is to promote the feature fusion of images and tooth contours
- A novel attention module is proposed to localize discriminative parts
- An improved loss is proposed to address easy samples and hard samples

Authors

Hu Chen, Che Sun, Peixi Liao, ..., Yi Lin, Zhenhua Deng, Yi Zhang

Correspondence

fanfei@scu.edu.cn (F.F.), yzhang@scu.edu.cn (Y.Z.)

In brief

Panoramic dental images play a significant role in identifying unknown bodies. While tooth contours are significant in classical methods, few studies using deep learning methods devise an architecture specifically to introduce tooth contours into their human identification models. Our model was tested on a large dataset consisting of 23,715 panoramic X-ray dental images with tooth masks from 10,113 patients, achieving an average rank-1 accuracy of 88.62% and rank-10 accuracy of 96.16%, which is much higher than other models.

Chen et al., 2022, Patterns *3*, 100485
May 13, 2022
https://doi.org/10.1016/j.patter.2022.100485

Patterns

CellPress
OPEN ACCESS

Article

A fine-grained network for human identification using panoramic dental images

Hu Chen,[1] Che Sun,[1] Peixi Liao,[3] Yancun Lai,[1] Fei Fan,[2,*] Yi Lin,[1] Zhenhua Deng,[2] and Yi Zhang[1,4,*]
[1]College of Computer Science, Sichuan University, Chengdu, Sichuan, China
[2]West China School of Basic Medical Sciences & Forensic Medicine, Sichuan University, Chengdu, Sichuan, China
[3]Department of Scientific Research and Education, The Sixth People's Hospital of Chengdu, Chengdu, Sichuan, China
[4]Lead contact
*Correspondence: fanfei@scu.edu.cn (F.F.), yzhang@scu.edu.cn (Y.Z.)
https://doi.org/10.1016/j.patter.2022.100485

THE BIGGER PICTURE DNA, fingerprints, faces, etc. have been used in human identification, but they are susceptible to decay when people die. Teeth do not decay, so experts use teeth as an effective feature in individual identification. In earlier times, experts did the comparation manually. Our model contains a branch devised specially to extract tooth contour features, which have proved to be meaningful in previous methods. With other improvements added, our model is able to identify the target person in 1,000 X-ray dental images with an accuracy of 88.62. There also exist limitations. The proposed model rests on masks, so in subsequent studies, we will perform unsupervised methods on teeth or other structures.
Compared with DNA, panoramic dental X-ray images are easier to access, so our model provides a feasible approach for identifying unknown bodies if they took panoramic dental X-ray images when alive, even if these bodies are ossified.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

When accidents occur, panoramic dental images play a significant role in identifying unknown bodies. In recent years, deep neural networks have been applied to address this task. However, while tooth contours are significant in classical methods, few studies using deep learning methods devise an architecture specifically to introduce tooth contours into their models. Since fine-grained image identification aims to distinguish subordinate categories by specific parts, we devise a fine-grained human identification model that leverages the distribution of tooth masks to distinguish different individuals with local and subtle differences in their teeth. First, a bilateral branched architecture is designed, of which one branch was designed as the image feature extractor, while the other was the mask feature extractor. In this step, the mask feature interacts with the extracted image feature to perform elementwise reweighting. Additionally, an improved attention mechanism was used to make our model concentrate more on informative positions. Furthermore, we improved the ArcFace loss by adding a learnable parameter to increase the loss of those hard samples, thereby exploiting the potential of our loss function. Our model was tested on a large dataset consisting of 23,715 panoramic X-ray dental images with tooth masks from 10,113 patients, achieving an average rank-1 accuracy of 88.62% and rank-10 accuracy of 96.16%.

INTRODUCTION

The identification of individuals plays an important role in forensics. Several biometric features, including DNA, fingerprints, faces, and voices, have been used in human identification. However, these features are susceptible to decay in the natural environment. In contrast, teeth are covered by enamel, which is the hardest tissue in the human body. Teeth do not undergo major morphological changes after death, which makes teeth effective features in matching antemortem and postmortem records.

Digital images are commonly used in human identification due to their acceptable costs and accuracy. Early on, antemortem and postmortem records were manually compared by experts using dental restoration and tooth morphology,[1,2] which incurred

significant time costs with subjective errors. Then, automated comparison systems were developed, and classical methods were adopted. Jain and Chen[3] used a semiautomatic contour extraction method to address the problem of fuzzy tooth contours and then retrieved the best match to a given postmortem radiograph from a database of 100 images. Nomir and Abdel-Mottaleb[4] used integral projection to separate teeth, and the signatures of vectors of tooth contours were compared. In recent years, deep neural networks have been applied in this area and have made substantial progress. Fan et al.[5] built a human identification system with a convolutional neural network using panoramic dental radiographs. Lai et al.[6] presented a network that incorporated a channel attention module to selectively emphasize channel information. Sathya and Neelaveni[7] used transfer learning and classified teeth into four classes, thereby increasing the performance of the model. Wu et al.[8] proposed a multisupervision network with an attention-based mechanism to obtain an attention mask for human identification. These previous studies substantially improved the accuracy and reliability of human identification using dental images. However, these deep-learning-based works used entire images as input and processed the textural feature together with the contour feature without distinction. Tooth contours have not been treated in much detail but have been proved to be meaningful features in previous methods.[9–11]

Fine-grained image classification refers to problems of distinguishing subordinate categories among basic categories.[12] In this paper, we treat the human identification problem as a fine-grained classification problem since the targets are the same. Moreover, fine-grained classification requires the model to focus on areas with different details but that are similar overall; in our model, we adopted several methods to distinguish fine-grained categories.

First, our model is composed of a basic branch and a mask branch to pay strong attention at the part level. The basic branch extracts the contextual deep visual features from the whole image, while the mask branch processes the masks (i.e., segmented tooth contours) to obtain the mask features. The mask features cooperate with the contextual features to generate the weighted features. The mask features can be seen as weights of the basic features, similar to convolutional filters in dynamic filter networks.[13] They enhance the correlation between bilateral networks, guide the model to pay strong attention to the fine-grained tooth contours, and adjust the training procedure automatically. Second, as a supplement to the strong attention mentioned above, to learn discriminative parts inside and outside of masks automatically, an improved attention mechanism is contained in the basic branch. The mechanism can also ease the constraint of the local receptive field brought by the convolution operation. Furthermore, to distinguish similar individuals (hard samples), inspired by LCANet,[6] we improved the ArcFace loss[14] and propose an improved loss function (dynamic ArcFace loss) as the loss function of the basic features and weight features.

The main contributions of our work are summarized as follows: (1) we propose a bilateral network model based on masks and feature fusion to solve the individual identification problem using dental panoramic X-ray images. The network architecture of this model is uncomplicated, and a large number of experiments have proved that it achieves better results than previous methods. (2) An attention module is devised that is simple but effective and can be easily used in other existing models without modifying their backbones. (3) We present an improved loss function (the dynamic ArcFace loss) that adjusts the loss in the training process using learnable parameters and generates more discriminative features. Our conclusions are confirmed by further experiments. On the test set, which is composed of 1,000 individuals, our model achieves a rank-1 accuracy of 87.81% and a rank-10 accuracy of 96.67%, and the accuracy is 82.92% when the false acceptance rate (FAR) is one in 10,000. The average rank-1 and rank-10 accuracy rates of the 5-fold cross-validation experiment are 88.62% and 96.16%, respectively.

RESULTS AND DISCUSSION

Related work

Fine-grained models

Fine-grained classification models are designed to distinguish between multiple subcategories, among which the differences can be subtle. Part-based R-CNNs[15] extract features from whole-object and part detectors and handle classification issues using methods borrowed from object detection. Wei et al.[16] proposed a mask-CNN model for aggregating contextual and part attributes based on part-annotated fine-grained images. Yang et al.[17] proposed a self-supervision mechanism consisting of a navigator network, a teacher network, and a scrutinizer network to localize informative regions automatically while optimizing these networks in a pipeline. Our bilateral network infuses information derived from the tooth contour mask into the main contextual feature and trains branches together.

Attention modules

By using attention modules, networks are able to determine long-range dependencies and examine more meaningful areas. The global attention mechanism proposed in the nonlocal network[18] contains a self-attention module that models multilevel global dependency relationships. GCNet[19] soon followed. GCNet combines the channel attention mechanism (SE block) in SENet[20] and the global attention mechanism in the nonlocal network[18] to propose a global attention mechanism global context (GC) block that derives contextual information to achieve global dependency modeling. Built on the SE block, the channel attention module in LCANet introduces a structure composed of BN-PreLU-Conv1*1-Dropout (LCANet module), which keeps the channel interaction in a proper range and promotes convergence. In our model, an LCANet model is used in the lateral branch of the GC block, resulting in an enhanced learning capability.

Cosine loss function

Solving the human identification problem on a large dataset deeply relies on devising a proper loss function. Early biometric vision tasks used the softmax loss as the loss function of the model. Subsequently, improved loss functions based on the softmax loss function to better distinguish features have been proposed. Among these improvements, Wen et al.[21] proposed the center loss to learn the intermediate features of each individual to reduce the intraclass gap, and Kemelmacher-Shlizerman et al.[22] proposed the L-softmax loss by adding angular constraints to the features to improve the discernibility of features.

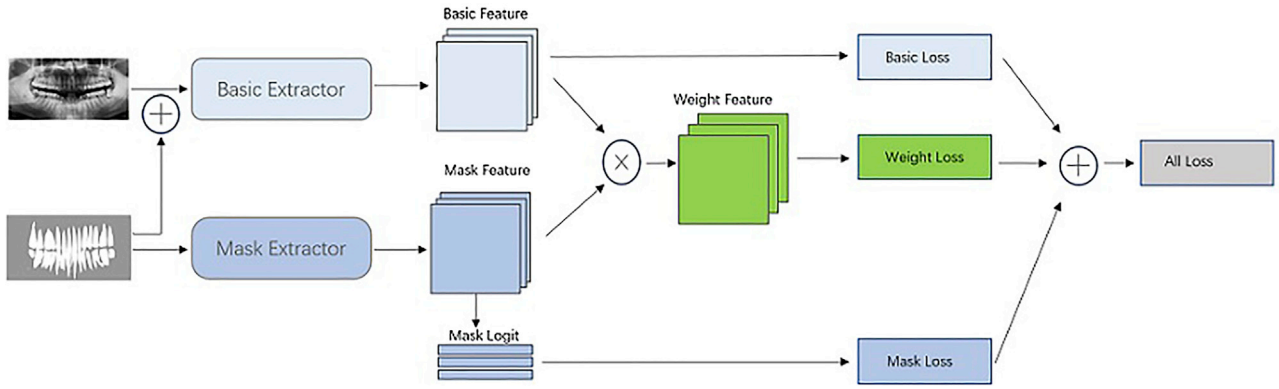


Figure 1. An overview of our network

Subsequently, Liu et al.[23] improved the accuracy of the L-softmax loss on the open set recognition task by normalizing the weight parameters and proposed the A-softmax loss. Wang et al.[24] proposed a cosine loss to improve the softmax loss. The cosine loss adjusted the features and weights through L2 normalization and added angular constraints to features. Then, the ArcFace loss proposed by Deng et al.[14] improved the cosine loss and changed the angular constraint of the feature from the cosine value to the angular value comparison so as to reduce the intraclass gap and expand the interclass gap. Our improved loss function adjusts the weights automatically to better learn the distribution of hard samples.

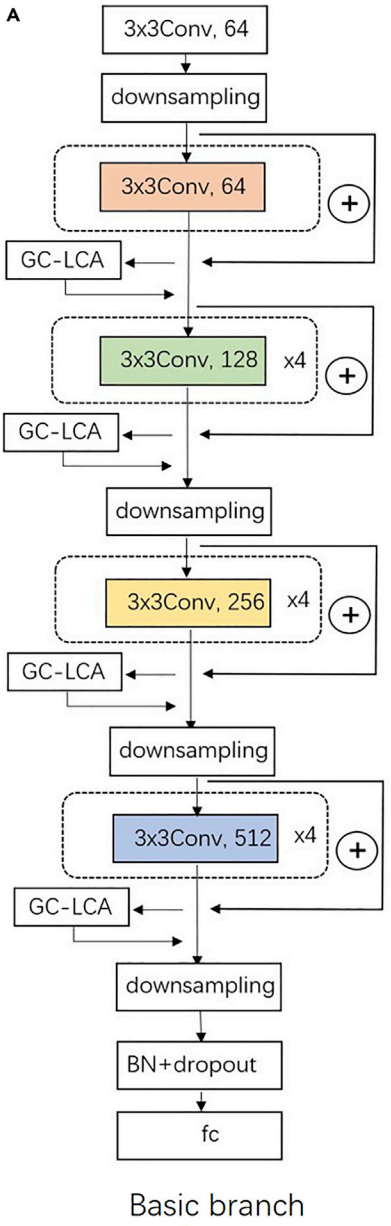
Methods

Bilateral-branch network

In this part, we illustrate a bilateral-branch network model, which pays strong attention to tooth contours. The basic branch extracts the basic features, and the mask branch extracts mask features to enhance the identification from a more fine-grained perspective. The framework is shown in Figure 1.

Basic branch. To excavate the features concerning tooth morphology from the mask, we combine the original tooth image and its corresponding tooth segmentation map (i.e., mask) using the channel axis and use the combination as the input of the basic network. Then, the model extracts the contextual features from the original image and local tooth contour features from the mask. Note that unless otherwise specified, the convolutional layer in the basic branch is composed of a normal convolutional layer, batch normalization (BN), and an activation function (parametric rectified linear unit [PreLU][25]). After being sent to the network, the input first goes through a convolutional layer with a 3 × 3 convolutional kernel and 64 channels and then is downsampled by maximum pooling. Next, it goes through four convolutional modules (I, II, III, IV). Each module is composed of BN and two convolutional blocks (each convolutional block is composed of two convolution layers with a 3 × 3 kernel size), and the numbers of channels of the blocks are 64, 128, 256, and 512 in sequence. In addition, after the convolution modules (II, III, IV), a convolutional layer with a step size of 2 and a 1 × 1 convolutional kernel is added to complete the downsampling operation (only a normal convolutional layer and BN are used in the downsampling convolutional layer). Due to the local receptive field of the convolutional neural network, the deeper the network is, the more features fade. To alleviate the harm caused by the local receptive field, two methods were adopted to enhance feature propagation. First, each convolutional block in the above four convolution modules adopts a head-to-tail feature fusion strategy to enhance feature propagation and reuse parameters. In addition, at the end of each convolutional block, an improved attention mechanism is included. It models the global contextual dependencies in the feature map better and strengthens the ability to obtain key information in the feature map. Subsequently, a structure composed of BN and dropout is added at the end of all convolutional layers to adjust the distribution of features. Finally, a fully connected layer is used to extract the 1,024-dimensional deep visual features of teeth. The flowchart is shown in Figure 2A.

Mask branch. Due to the importance of tooth contours in classical human identification methods,[3,4] we devise a specific mask branch to make full use of the morphological characteristics of teeth. Given a tooth segmentation image (i.e., mask) determined by convolutional neural network (CNN) corresponding to a panoramic dental image, there are only two different pixel values in the segmentation image: a value of one for the pixels in the target area of the tooth, and a value of zero for the background area. That binary distribution strengthens the mask branch such that it can capture the slight feature distribution changes in different masks accurately. After passing to the mask branch, the mask image goes through four convolutional blocks, and each convolutional block is composed of two convolutional layers with a 3 × 3 convolutional kernel. Each convolutional layer is composed of a normal convolutional layer, a group normalization (GN) layer and an activation function (PReLU) structure. The numbers of channels of the four convolutional blocks are 64, 128, 256, and 512, respectively. A 1 × 1 convolutional kernel with a step size of 2 is set after each convolution block to downsample the feature map. Finally, two fully connected layers are used: one obtains the 1,024-dimensional mask features and interacts with the basic features, and the other predicts the class (mask logit) and calculates the mask loss directly. Our branch network replaces traditional BN with GN and utilizes the trait that GN adjusts the data distribution within the channel group to improve the characterization capabilities of mask features. The flowchart is shown in Figure 2B.



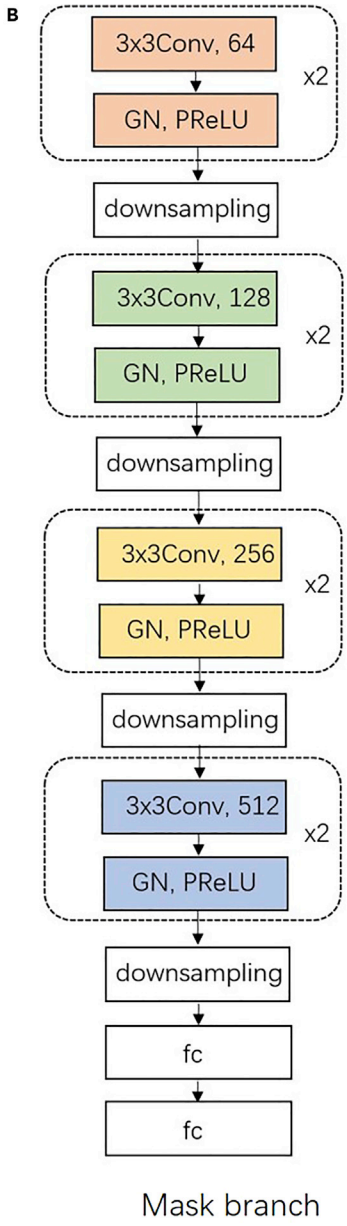


Figure 2. Flowcharts

(A) Model architecture in basic branch. (B) Model architecture in mask branch.

Feature fusion. In dental panoramic X-ray images, the background area is much larger than the target area; additionally, due to different formal clinical treatments, many individuals have similar tooth characteristics. As shown in Figure 3, different individuals have lost the same teeth and have had new teeth implanted, which easily leads to a higher similarity of the characteristics excavated by the model. In the last subsection, we devised a mask branch that extracts the mask features to better collect the morphological information of teeth. Until now, our branch networks have been two separate units. To strengthen the connection between the two branches, motivated by Kumar et al.,[26] we adopt an elementwise feature fusion strategy. This strategy performs elementwise multiplication on the features extracted by the basic branch and mask branch. Furthermore, the human identification problem can be treated as a fine-grained classification problem with small samples. This feature fusion strategy can help the network discover more discriminative visual features from tooth images in each subclass (one person) and overcomes the existing challenging small-sample problem. The practical significance of each part of the model will be proved by ablation studies in the experimental section.

GC-BN-PReLu-Conv1x1-Dropout attention module

The attention mechanism has become a feasible approach to fine-grained image recognition.[27] Since directly repeating convolution layers would lead to ineffective modeling of the long-range dependency,[19] GCNet proposes the GC block (as shown in Figure 4C) to model the global context. Specifically, the GC block utilizes a context modeling module (as shown in the red box of Figure 4C) which contains a Conv 1*1 and a matrix multiplication to form a global context feature.[19] However, the GC block projects channel features into a one-dimensional space, which may omit information in other dimensions[28–30] and cause weak global interactions among the global context features in the human identification task with dental images. As shown in Figure 5B, we visualized the attention map (heatmap) of the GC block for different positions. It was observed that the responses in the surrounding background and the target tooth region were analogous and salient, indicating that global context features are indiscriminative for different positions. To statistically verify this observation, we analyzed the SD value of attention weights in different positions in Table 1) and determined that the value is small, indicating that attention weights learned by the GC block are similar for different positions, which further verifies that the GC block is inefficient in modeling the long-range dependencies.

Inspired by LCANet,[6] in which the BN-PReLu-Conv1x1-Dropout (BPCD) block can capture interactions, we attempt to develop an efficient attention instantiation by developing the BN-PReLu-Conv1x1-Dropout module at the context modeling module to aggregate the global context features and reach a new instantiation of the general framework, which is referred to as the GC-BPCD block. Specifically, the BPCD module can be regarded as vector concatenation-based attention.[6] Compared with the original context modeling module (as shown in Figure 4C), the

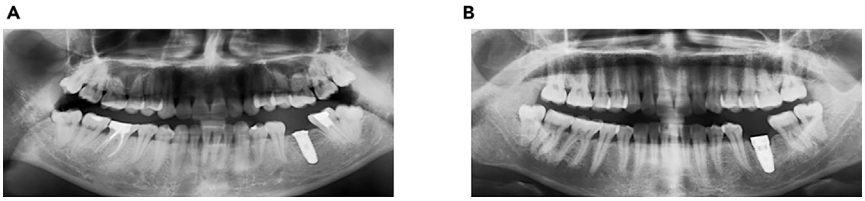


Figure 3. Similar radiographs of two different individuals, who have lost the same teeth at the same place and had new ones implanted

BPCD module can strengthen the globe interactions and feature aggregation by repeatedly refining the global context features, producing more effective long-range dependencies. Specifically, the coefficient of the negative part is not constant, which can improve the model fitting with minimal overfitting risk.

Our module is formulated as

$$z_i = x_i + W_2 RELU\left(LN\left(W_1 \sum_{j=1}^{N_p} \frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}} f(x_j)\right)\right), \quad \text{(Equation 1)}$$

where $f(\cdot)$ denotes

$$f(x) = BN(PReLu(Wx)). \quad \text{(Equation 2)}$$

For Equation 1, $x_i = \{x_i\}^{N_p}$ is the input feature, where $N_p = H \cdot W$ is the number of positions in the feature map. W_k represents the parameters in Conv 1x1 before softmax; $f(x_j)$ denotes LCA module(Equation 2); and $\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}}$ denotes softmax. $\sum_{j=1}^{N_p} \frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}} f(x_j)$ corresponds to the context modeling part in Figure 4D, and the remaining $W_2 RELU(LN(W_1(\cdot)))$ denotes the transform part in Figure 4D, where W_1 and W_2 are parameters of two Conv 1×1 s. $x_i + (\cdot)$ denotes the residual part.

For Equation 2, equation denotes the BPCD part in Figure 4B, and W is the parameter of Conv 1x1.

To prove the effect of our attention module, we choose three images randomly and then calculate the SD of the channel attention values in the GC-BPCD module and GC block separately. The results of the proposed model are shown in Figure 5C and Table 1. The responses between the surrounding background and the target tooth region are different and the SD value is higher, which indicates that the proposed GC-BPCD block strengthens the modeling of the long-range dependencies and enables better recognition performance. More experiments are discussed later in the experiment section.

Dynamic ArcFace loss

In the last two subsections, part-level strong attention and soft attention have been established to focus on discriminative parts, and then to promote fine-grained feature learning. In this subsection, an improved loss is devised to improve the discernibility of features at the object level. For a better comprehension of our dynamic ArcFace loss, in this section, we first review the design of the ArcFace loss and former loss measures.

Revisiting ArcFace loss and the former. The softmax loss, known as the most classical loss function in biometric recognition, can be expressed as

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}, \quad \text{(Equation 3)}$$

where x_i is the feature of the ith sample of the y_ith category and has a dimension of d, $W \in \mathbb{R}^{d \times n}$ denotes the weight parameter, W_j denotes the jth column of W, $b_j \in \mathbb{R}^n$ is the bias, N is the batch size, and n is number of categories.

Despite its simplicity, the softmax loss fails to help the model extract unique features. The following cosine loss performs the corresponding improvement, and it is expressed as

$$L_2 = -\frac{1}{N}\sum_{i} \log \frac{e^{s\left(\cos\left(\theta_{y_i,i}\right)-m\right)}}{e^{s\left(\cos\left(\theta_{y_i,i}\right)-m\right)} + \sum_{j \neq y_i} e^{s\cos\left(\theta_{j,i}\right)}}, \quad \text{(Equation 4)}$$

$$W = \frac{W^*}{\|W^*\|}, x = \frac{x^*}{\|x^*\|}, \cos(\theta_j, i) = W_j^T x_i,$$

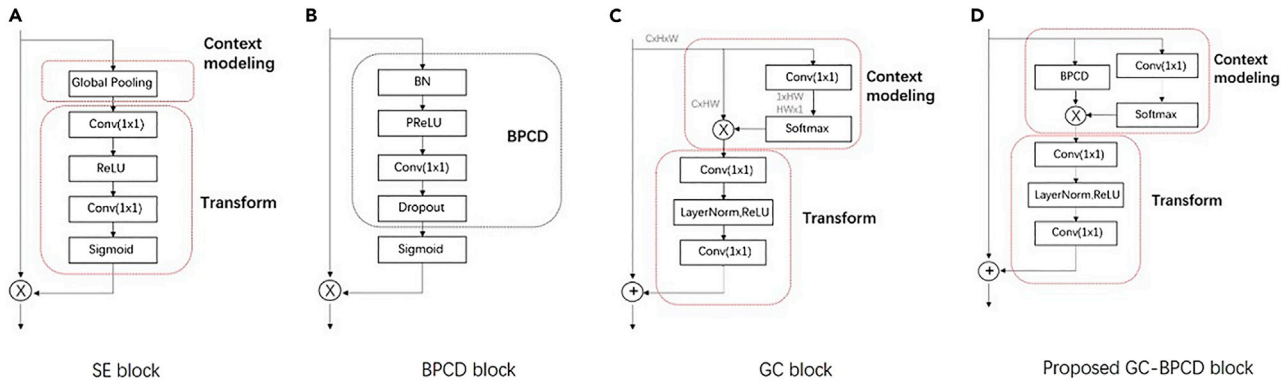


Figure 4. Architecture of channel attention blocks

(A) SE block. (B) BPCD block. (C) GC block. (D) Proposed GC-BPCD block.

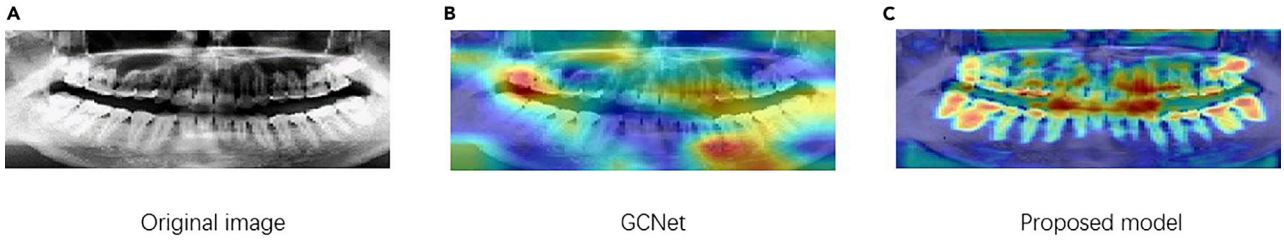


Figure 5. Heatmaps

(A) is the original image, and (B) and (C) are heatmaps generated by different models using (A). (B) is generated by GCNet, (C) is generated by our poposed model.

where θ_j is the angle between W_j and x_i. The cosine loss adjusts feature x_i and weight W_j using L_2 normalization and adds a restriction m on the angle.

Subsequently, the ArcFace loss further improves the cosine loss, and the restriction on the cosine is directly changed to restriction on the angle as

$$L_3 = -\frac{1}{N}\sum_i \log \frac{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}+\sum_{j=1,j\neq y_i} e^{s\cos\theta_j}}, \quad \text{(Equation 5)}$$

This increases the interclass distance and decreases the intraclass distance. Taking a binary classification problem as an example, θ_i represents the angle between weight W_i and feature x_i in category $C_i (i = 0, 1)$. Assuming that a certain feature belongs to category 0, then the ultimate goal of the cosine loss is to make $cos(\theta_0) - m > cos(\theta_1)$. The cosine function is monotonically decreasing in the range of $(0, \pi)$, so increasing the angular constraint m forces the model to learn from the features that have larger interclass distances and smaller intraclass distances. Similarly, the ArcFace loss applies constraints directly to the angular value in order to make $cos(\theta_0 + m) > cos(\theta_1)$, which further increases the interclass distance and decreases the intraclass distance. The ArcFace loss experiment proves that the ArcFace loss improves the performance and is better than the cosine loss. However, as shown in Figure 7, our experiment reveals that there is a definite increase in the ArcFace loss that makes the training procedure more difficult, and the loss remains basically unchanged at approximately 0.1 in the late stage of training. Huang et al.[31] show that the ArcFace loss does not pay enough attention to difficult samples, which relatively increases the loss from simple samples, thus stabilizing the loss value since the model has fully learned the distribution of simple samples.

Dynamic ArcFace loss (improved loss). To solve this problem, inspired by focal loss,[32] we seek to reduce the loss produced by those easy samples. The formula of our loss function is expressed as

$$L_4 = \left(1 - \frac{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}+\sum_{j=1,j\neq y_i} e^{s\cos\theta_j}}\right)^2 \times L_3, \quad \text{(Equation 6)}$$

Table 1. SD values

Module	GC block	Proposed model
Images	7.12	11.18

where L_3 is the ArcFace loss (defined in Equation 5).

In the training procedure, our dynamic ArcFace loss suppresses easy samples with large losses while promoting hard samples with small losses by increasing their loss values, thus encouraging the model to learn hard samples automatically.

Different from curricular loss, the proposed dynamic ArcFace loss adjusts the weights of samples outside of the log function and is iteration agnostic. Different from focal loss, we introduce cosine similarity for human identification tasks and set γ to 2 under the objective of robustness. Different from ArcFace loss, dynamic ArcFace loss is equipped with a flexible mechanism to address the difficulty of samples. All parts of the proposed loss function are devised to repair the former drawbacks. The performance of dynamic ArcFace loss also outperforms ArcFace loss and curricular loss, as Figure 7 shows.

The final loss is an aggregation of the dynamic ArcFace loss and L_1 loss. We use the dynamic ArcFace loss in the optimization of basic features L_{basic} and weight features L_{weight}. For mask loss, since the mask logit in Figure 1 is a direct prediction of individuals, we choose the L_1 loss. The final loss can be expressed as

$$L_{final} = L_{basic} + L_{weight} + \lambda L_{mask} \quad \text{(Equation 7)}$$

where λ is the hyperparameter and is set to 0.1. This is done because, at the beginning of the training procedure, L_{mask} fluctuates extremely and, if it is set to 1, the model will not converge. More improvements will be given in the section "experimental procedures."

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yi Zhang (yzhang@scu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- We have produced and used human data (dental images) in this study, so, due to ethical consideration, we are not able to release the dataset publicly. However, interested researchers can ask the lead contact for access to the data. Codes for panoramic dental images are available at https://github.com/BreezeHavana/FGHINet.
- All original code has been deposited at Zenodo under the DOI 10.5281/zenodo.6223257 and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

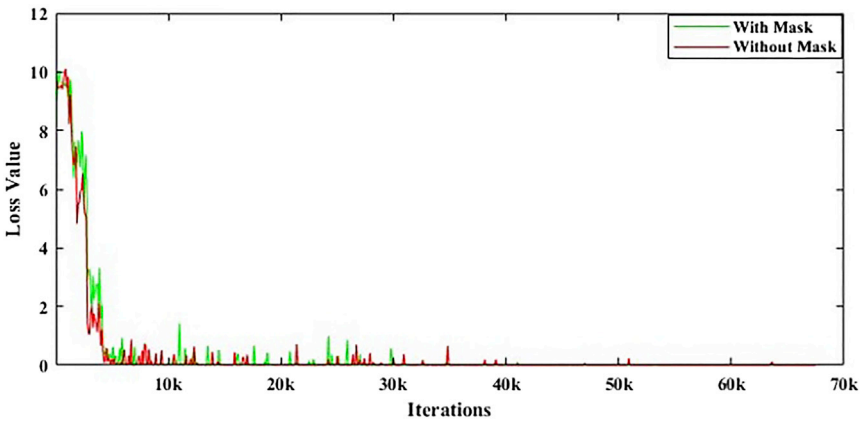


Figure 6. Softmax loss curve
Best viewed in color.

Ethics statement

All study participants provided informed consent, and the study design was approved by an ethics review board.

Data acquisition and hyperparameters

All the panoramic X-ray dental images used in our experiments, which were taken before and after oral clinical treatment, were collected from the West China Hospital of Stomatology and the Peking University Hospital of Stomatology. The dataset contains tooth filling, root canal, and restorations. Images derived from different devices do not affect the results. Our dataset consists of 23,715 images from 10,113 patients. The dataset is randomly split into a training set that consists of 21,575 images from 9,113 patients and a test dataset composed of 2,140 images from 1,000 patients. The training set and test set are completely separated from the individual axis; that is, if an image belonging to an individual is chosen by the training set, all the images of the individual should be collected in the training set, which is a concept borrowed from human identification.[22] For better generalization performance, we further separate the test set into a register set and a validation set. If there are M total dental images that belong to an individual, we randomly pick one for the register set, and the other $M - 1$ images are the validation set.

Our device is equipped with Nvidia Geforce 2080Ti; the cuda version is 10.2. A total of 9 s is needed to train 100 iterations, and the training time is approximately 2 h. The inference time is 0.01 s per image. For preprocessing, We first crop the image size to (2,100, 850) to exclude irrelevant areas and resize the scale from (2,100, 850) to (128, 128) as the size of input image, then we reduce the influence of artifacts in the original image using contrast enhancement, and finally we set rectangles with a size of (15, 30) in the target area randomly to perform random occlusion. This occlusion is performed to improve generalization and stability of the model. The masks in our dataset are determined by CNN. We choose stochastic gradient descent (SGD) with a momentum of 0.9 and a decay of 0.0005 as our optimizer. The batch size is 16, the dropout is 0.6, and the initial learning rate is 0.01. The learning rate decreased by half every 30,000 steps, and all training steps are completed after 80,000 steps. In the testing procedure, we only use the basic branch to extract 1,024-dimensional features, use the cosine distance as the similarity, and then rank them. More specifically, we separately extract the features of images from the register set and validation set by a trained model, choose an image from the validation set randomly, match its feature with the feature of every image in the register set, and calculate their similarities. Then, we rank all the similarities from high to low. Finally, the identity of the image is determined by the image in the register set with the highest similarity.

Ablation studies in basic branch

In this section, we perform several ablation studies to prove the importance of each part of the network. All the experiments are implemented without a mask branch or feature fusion in this part.

Masks

Concentrating on the role masks play in our model, we abandoned the other modified units mentioned above with only the basic backbone and softmax loss left and used few images and a combination of images and masks as input, respectively. As shown in Table 2, after we added masks as supplemental information, the rank-1 accuracy increases by 4.47% and the true acceptance rate (TAR) increases by 4.83% when FAR is 1×10^{-4}. The results prove that adding masks improves the performance, and the accuracy is still low. Moreover, we visualize the training loss in Figure 6, which indicates that our model converges fast regardless of whether masks are added, and the main reason is that the softmax loss is easy to optimize and converges fast.

Dynamic ArcFace loss (improved loss)

In this section, we continue to conduct experiments on the impact of our improved loss, and the combination of images and masks is taken as the input of the basic branch. Five other loss functions (L-softmax loss, A-softmax loss, cosine loss, ArcFace loss, and curricular loss[31]) are adopted as comparisons to verify the improved loss function in this chapter. As shown in Table 3, among all loss functions, the cosine loss, ArcFace loss, and dynamic ArcFace loss have better results than other loss functions because these three losses add

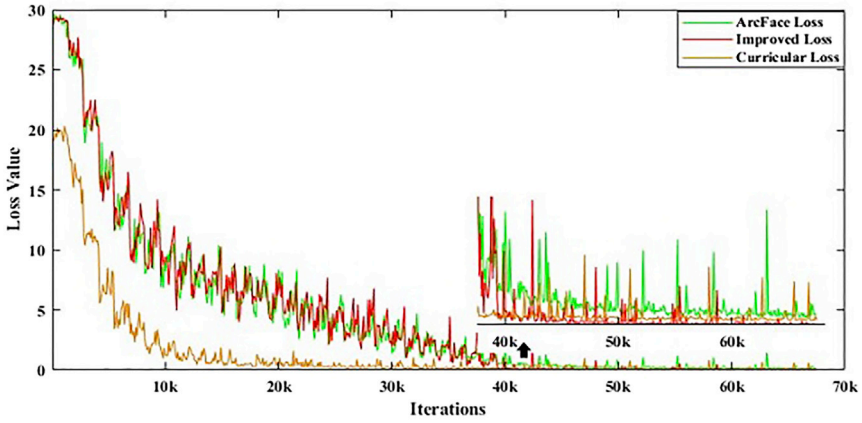


Figure 7. Loss curves of three different loss functions
Best viewed in color.

Table 2. With or without masks

Input	Rank-1 accuracy	Rank-10 accuracy	TAR(@FAR = 10-4)
Images	34.47	57.63	26.39
Images + masks	38.94	63.52	31.22

angular constraints to features, and it is beneficial for the model to learn features with larger interclass distances and smaller intraclass distances. Furthermore, the ArcFace loss and dynamic ArcFace loss directly affect the angle, thus achieving better results than the cosine loss. In addition, both the curricular loss and the dynamic ArcFace loss improve the ArcFace loss. The difference is that, in different training procedures, the angular constraint weights are determined by different numbers of iterations, and the weight increases as the iteration number increases, which leads to a fast decrease in the loss. However, our dynamic ArcFace loss adjusts the weight by the similarity score of the sample to reduce the loss of easily distinguishable samples, which is why the dynamic ArcFace loss performs better than the curricular loss. We also visualize the training losses of different loss functions, as shown in Figure 7. Figure 7 shows that, when the ArcFace loss is used in training, the loss curve shows a significant decreasing tendency, which converges in the later stage, but the final loss value is still high, indicating that there is still room for optimization. Our dynamic ArcFace loss function not only maintains the convergence of the original ArcFace loss but also pays more attention to hard samples, thus reducing the final loss value of the model and further proving that its self-learning weight can adjust the loss value in the network training process. The curve of the curricular loss accelerates the convergence of the model (the model converges after 30,000 iterations), but it does not reduce the loss value as much as the dynamic ArcFace loss. It can be inferred that the curricular loss has limited optimization effects on the model. In general, the experiments above prove the superiority of the dynamic ArcFace loss and prove the feasibility that the dynamic ArcFace loss can be used as the loss function of our model.

GC-BPCD module

To determine the effect the GC-BPCD module has on the basic branch, we performed several correlational experiments. We used the combination of the mask and the original image as the input of the basic branch and dynamic ArcFace loss as the loss function with the branch network and feature fusion excluded.

Table 4 reveals that, among the six modules embedded in the basic branch, the results of the LCANet module and GC-BPCD module exceed those of other modules, while the GC-BPCD module offers superior performance with a rank-1 accuracy of 75.53%. Compared with several other channel attention mechanisms (except the GC-BPCD module), the attention module in LCANet improves the ability of the channel attention mechanism in the original SENet to capture the linear dependency cross-channels and can selectively emphasize the key information that is effective for the recognition task. The global attention mechanism in GCNet has shortcomings in global contextual modeling and channel attention modeling. Our GC-BPCD module not only improves GCNet's ability to model the global contextual information in the feature map but also fully captures the channel dependency between any two channels.

Table 3. Different loss functions

Loss function	Rank-1 accuracy	Rank-10 accuracy	TAR(@FAR = 10-4)
L-softmax loss	44.94	67.52	38.22
A-softmax loss	38.57	63.25	27.11
Cosine loss	47.46	71.93	38.77
ArcFace loss	56.58	79.47	47.91
Curricular loss	45.79	68.60	38.19
Dynamic ArcFace loss	65.96	85.26	57.10

Table 4. Attention modules

Attention module	Rank-1 accuracy	Rank-10 accuracy	TAR(@FAR = 10-4)
SENet	51.14	75.53	41.15
CBAM	50.53	75.96	40.93
LCANet	68.60	89.04	60.14
ECA-Net	52.11	75.97	44.04
GCNet	64.74	85.96	57.63
Proposed module	75.53	90.87	67.82

We further select three different images randomly from the training set and use Grad-CAM for visualization. Figure 8 shows that the response areas obtained by the attention modules in SENet, CBAM, and ECA-Net are relatively scattered, while large-scale response areas appear in the target tooth area and the background area (such as SENet and CBAM), and these modules fail to exclude the interference caused by the background area, thus reducing the rank-1 accuracy. The heat produced by the GCNet module is not focused on target areas either, owing to the deficiencies the model has on contextual modeling and channel attention modeling. The LCANet module has improved the attention mechanism in SENet and focuses more attention on the tooth area. Our GC-BPCD module inherits this tendency, aggregates the contextual information of the target area, pays more attention to the target area, and pays less attention to the background.

The architecture of our GC-BPCD module is quite clear and simple. It compensates for the deficiencies of the global attention mechanism in the original GCNet and strengthens its ability to capture channel dependencies. The experiments above prove that the GC-BPCD module is a lightweight but effective unit in recognition tasks.

Ablation studies in mask branch

In this section, we conduct more experiments to evaluate our mask branch. The basic branch is added since the mask branch cannot work as a single model, and the parameters in the basic branch are set to be consistent with the final version.

Backbone

Table 5 shows the results of different backbones implemented in the mask branch without feature fusion. Among these results, U-Net[33] achieves better performance than ResNet,[34] probably because U-Net is designed specifically for segmentation tasks and derives mask features more precisely. This finding inspired us to design our mask branch model. Otherwise, the model yields better performance after we replace BN with GN, demonstrating that the representation ability of the mask branch has been strengthened. Since the mask branch uses a mask (which only contains two values) as input, GN adjusts the local feature distribution and makes features extracted from masks more discriminative.

Feature fusion

As shown in Table 6, after feature fusion is aggregated, a huge leap occurs in the accuracy, which implies that the power of the mask branch is limited if added directly. When we use elementwise multiplication instead of elementwise addition, the rank-1 accuracy increases by 6.67%. The elementwise addition operation adds redundancy while multiplication emphasizes features selectively. Weight features tighten the relationship between basic features and mask features, aggregate them, and optimize features together. They adjust the training procedure automatically and can be seen as weights of basic features, similar to convolution filters in dynamic filter networks.[13]

Loss function

The different choices of loss functions for the mask loss, including the L1 loss, softmax loss, and our dynamic ArcFace loss (improved loss), are given in Table 7, while the basic loss and weight loss are fixed as the dynamic ArcFace loss. As shown in Table 7, the accuracy of the softmax loss is lower than the accuracies of the other methods, while the L1 loss and dynamic ArcFace loss achieve comparable performance. Moreover, the type of loss function adopted in the mask loss affects the optimization of

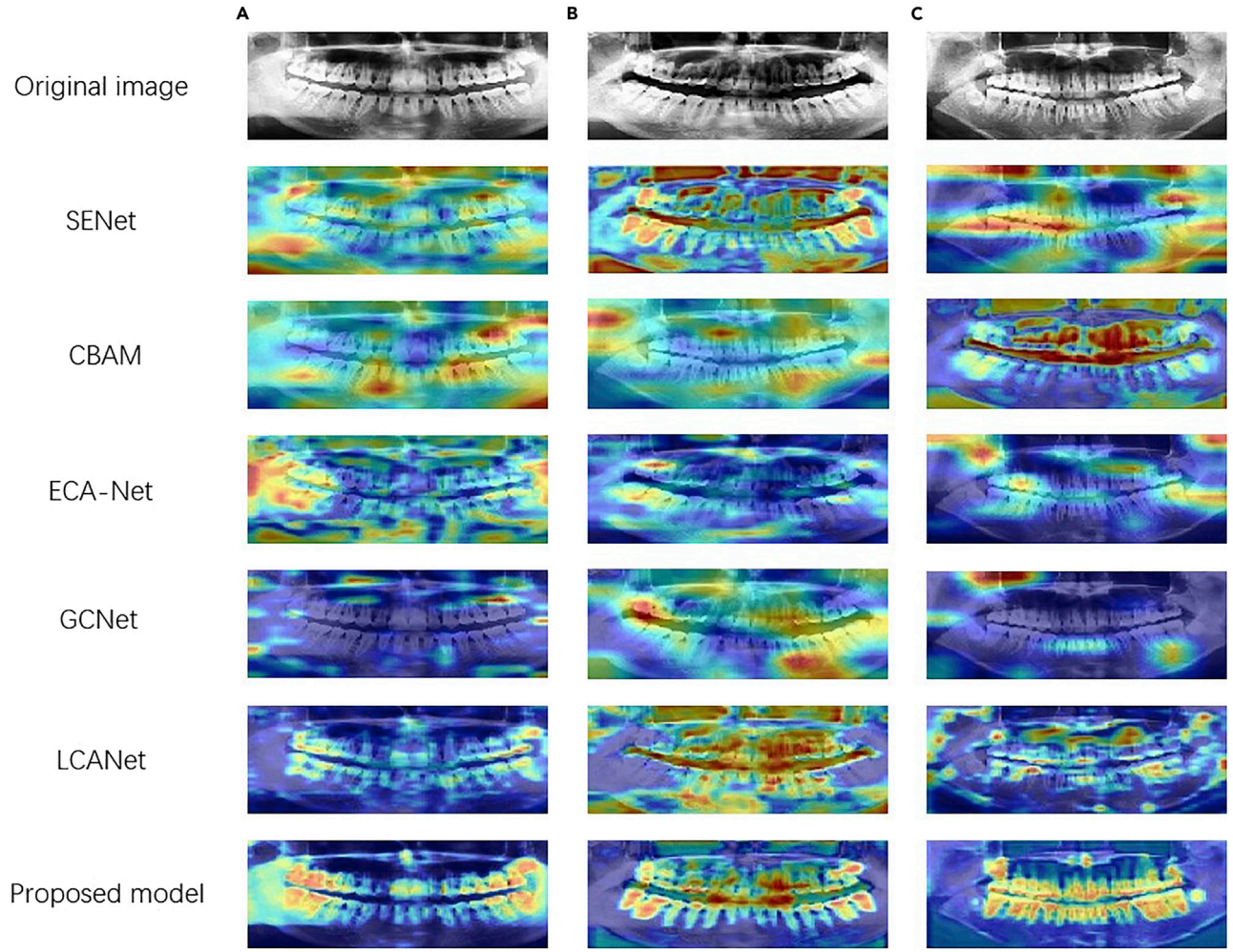


Figure 8. Grad-CAM visualizations

The first row contains three different images, and in the following rows, each row contains three heatmaps correspond to different images.

the basic branch. We visualize basic loss in Figure 9. The figure shows that the convergence trends for all the loss functions are the same, and the value of the L1 loss is lower. Additionally, the loss curve is smoother than that in Figure 7, indicating that the training procedure becomes easier.

Other classification models

We then evaluate several other classification or identification models on our dataset, and our model obtains the top accuracy. The results are shown in Table 8. First, we compare our model with classical classification models, including ResNet, ResNeXt,[35] InceptionNet,[36] DenseNet,[37] and EfficientNet.[38] Among these classical models, EfficientNet outperforms other models, and the performance of InceptionNet and DenseNet is not satisfactory. The reason is that InceptionNet and DenseNet are deeper than the others. DenseNet reuses parameters by concatenating channels in each dense block but loses information when 1 × 1 convolutions are used, and EfficientNet is embedded with the attention model in SENet, thus concentrating more on tooth areas.

Moreover, compared with the other four identification networks, our model performs well. Ajaz and Kathirvelu[39] and Oktay[40] are two classical methods applied on small datasets consisting of 200 images and have poor effects. DenseNet is a deep learning model composed of a backbone network similar to ResNet without dense connections, and its performance is not as good as those of LCANet and our network. LCANet aggregates self-learning weight features and attention mechanisms and has a better effect, but it does not make full use of dental morphological information.

Table 5. Backbones

Backbone module	Rank-1 accuracy	Rank-10 accuracy	TAR(@FAR = 10-4)
ResNet	56.92	71.83	48.52
U-Net	61.30	74.11	53.38
Proposed	67.62	77.53	62.34
Proposed + GN	69.82	87.01	61.50

Table 6. Feature fusion

Feature fusion	Rank-1 accuracy	Rank-10 accuracy	TAR(@FAR = 10-4)
No feature fusion	69.82	87.01	61.50
Elementwise addition	81.14	93.51	73.95
Elementwise multiplication	87.81	96.67	82.92

Table 7. Loss functions for mask loss

Loss function	Rank-1 accuracy	Rank-10 accuracy	TAR(@FAR = 10-4)
Softmax	83.16	94.65	78.06
Dynamic ArcFace loss	86.05	96.23	81.32
L1	87.81	96.67	82.92

Table 8. Other models

Model	Rank-1 accuracy	Rank-10 accuracy	TAR(@FAR = 10-4)
ResNet	54.21	77.72	42.97
ResNeXt	58.68	81.14	50.27
InceptionNet	48.77	75	39.03
DenseNet	44.56	71.05	34.32
EfficientNet	69.21	87.20	61.35
Ajaz and Kathirvelu	17.51	24.39	8.73
Oktay	26.75	47.37	18.00
DentNet	39.47	65.09	28.17
LCANet	78.86	92.81	72.67
Proposed	87.81	96.67	82.92

Cross-validation

To further evaluate our model, we perform 5-fold cross-validation, and the results are shown in Table 9. The final average rank-1 accuracy is 88.62%, the highest rank-1 accuracy is 89.91%, and the lowest is 87.74%. All these data indicate that our model performs well and has generalization ability.

Experimental conclusion

In the section "ablation studies in basic branch," we perform ablation studies in the basic branch. In the section "masks," we explore the role of masks. With other modified units abandoned, when masks are added, rank-1 accuracy increases from 34.47 to 38.94 by four points and is still low. We substitute softmax loss with other loss functions in the section "dynamic ArcFace loss (improved loss)." The dynamic ArcFace loss exceeds ArcFace loss by approximately nine points, which proves that it fully exploits the potential of the added masks. Next, we compare several channel attention modules. The proposed model outperforms the LCA module by approximately seven points, possibly because the proposed model fully captures the channel dependency between any two channels. A detailed analysis is provided in the section "GC-BPCD module." The rank-1 accuracy reaches 75.53 when only the basic branch is employed.

Ablation studies in the mask branch are performed in section "ablation studies in mask branch." First, in section "backbone," we compare backbone models (mask extractor in Figure 1). The accuracy of the proposed backbone model is 69.82, which outperforms other backbone models. There is no feature fusion (green part in Figure 1) at that stage, and the mask branch extracts features of masks and makes predictions in categories. Owing to the large number of classes, the model is not very accurate in its predictions, so the accuracy decreases to 69.82 from its value of 75.53 in section "ablation studies in basic branch." Next, in section "feature fusion," we perform different methods of feature fusion. Elementwise multiplication bridges the gap between images and masks and increases the accuracy by a large amount to 87.81. We compare different loss functions for mask loss and achieve a higher accuracy of L1 loss.

In section "other classification models," we evaluate other classification or identification models. In section "cross-validation," we perform 5-fold cross-validation and obtain an average rank-1 accuracy of 88.62.

The aforementioned ablation studies prove the effectiveness of every single part of our model. The usage of teeth masks is the highlight of the proposed model, but a simple combination just trails the desired accuracy. Directly concatenating masks with images (section "masks") or barely adding lateral connections (section "backbone") does not fully utilize masks, which limits the overall performance. Every part of the proposed model is devised to address the fusion problem. There also exist limitations. Teeth have an important role in the proposed model. The proposed model rests on masks, and extra time is needed to draw masks of teeth. In subsequent studies, we will perform unsupervised methods on teeth or other structures.

CONCLUSION

Formal works on human identification with panoramic dental images aim for classification but neglect the effect of dental morphology. Based on this limitation, we propose a novel fine-grained bilateral-branch network. Our key idea is to promote the feature fusion of images and masks, thus to help the model focus on meaningful tooth contours. A novel attention module (the GC-BPCD module) is proposed to localize discriminative parts, and an improved loss (dynamic ArcFace loss) is proposed to address easy samples and hard samples. Ablation studies prove that the proposed model outperforms other identification models devised specifically for the dataset and other general classification models.

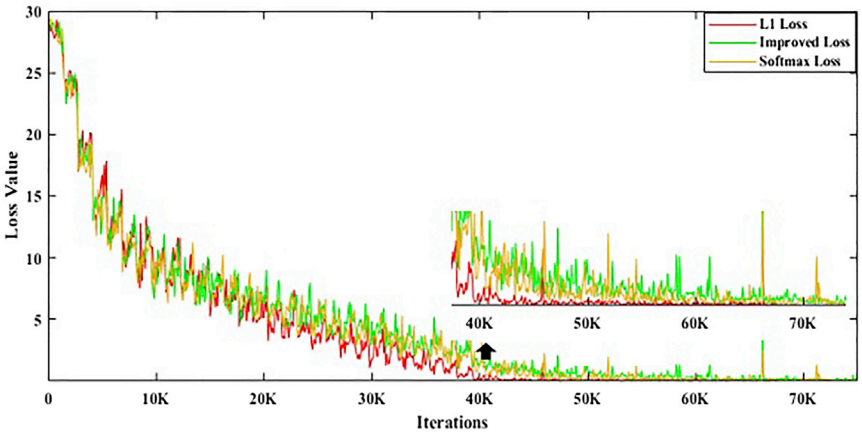


Figure 9. Loss curves of three different loss functions
Best viewed in color.

Table 9. Five-fold cross-validation

Number	Rank-1 accuracy	Rank-10 accuracy	TAR(@FAR = 10-4)
1	89.91	96.57	83.90
2	88.68	96.32	83.88
3	87.74	95.67	83.43
4	88.43	96.08	86.27
5	88.34	96.17	83.23
Average	88.62	96.16	84.14

Experiments elaborate that the proposed method is robust and easy to converge.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under grant 61871277 and grant 61671312, in part by the Sichuan Science and Technology Program under grant 2019YFH0193 and grant 2021JDJQ0024, and in part by the Chengdu Science and Technology Program under grant 2018YF0500069SN.

AUTHOR CONTRIBUTIONS

Conceptualization ideas, H.C.; methodology, H.C.; software programming, C.S., Y. Lai; visualization, C.S., Y. Lai; data curation, P.L., F.F., Z.D.; Writing – original draft, C.S., Y. Lai; Writing – review & editing, H.C., P.L., F.F., Z.D., Y.Z., Y. Lin; supervision, H.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 23, 2021
Revised: December 24, 2021
Accepted: March 8, 2022
Published: April 1, 2022

REFERENCES

1. Fahmy, G., Nassar, D., Haj-Said, E., Chen, H., Nomir, O., Zhou, J., Howell, R., Ammar, H.H., Abdel-Mottaleb, M., and Jain, A.K. (2004). Towards an automated dental identification system (adis). In International conference on biometric authentication (Springer), pp. 789–796.
2. Brogdon, B.G. (1998). The scope of forensic radiology. Clin. Lab. Med. *18*, 203–240.
3. Jain, A.K., and Chen, H. (2004). Matching of dental x-ray images for human identification. Pattern Recognition *37*, 1519–1532.
4. Nomir, O., and Abdel-Mottaleb, M. (2005). A system for human identification from x-ray dental radiographs. Pattern Recognition *38*, 1295–1305.
5. Fan, F., Ke, W., Wu, W., Tian, X., Lyu, T., Liu, Y., Liao, P., Dai, X., Chen, H., and Deng, Z. (2020). Automatic human identification from panoramic dental radiographs using the convolutional neural network. Forensic Sci. Int. *314*, 110416.
6. Lai, Y., Fan, F., Wu, Q., Ke, W., Liao, P., Deng, Z., Chen, H., and Zhang, Y. (2020). Lcanet: learnable connected attention network for human identification using dental images. IEEE Trans. Med. Imaging *40*, 905–915.
7. Sathya, B., and Neelaveni, R. (2020). Transfer learning based automatic human identification using dental traits-an aid to forensic odontology. J. Forensic Leg. Med. *76*, 102066.
8. Wu, Q., Fan, F., Liao, P., Lai, Y., Ke, W., Du, W., Chen, H., Deng, Z., and Zhang, Y. (2021). Human identification with dental panoramic images based on deep learning. Sensing Imaging *22*, 1–19.
9. Zhou, J., and Abdel-Mottaleb, M. (2005). A content-based system for human identification based on bitewing dental x-ray images. Pattern Recognition *38*, 2132–2142.
10. Nomir, O., and Abdel-Mottaleb, M. (2008). Hierarchical contour matching for dental x-ray radiographs. Pattern Recognition *41*, 130–138.
11. Lin, P.-L., Lai, Y.-H., and Huang, P.-W. (2012). Dental biometrics: human identification based on teeth and dental works in bitewing radiographs. Pattern Recognition *45*, 934–946.
12. Lin, T.-Y., RoyChowdhury, A., and Maji, S. (2015). Bilinear cnn models for fine-grained visual recognition. In Proceedings of the IEEE international conference on computer vision (IEEE), pp. 1449–1457.
13. Jia, X., De Brabandere, B., Tuytelaars, T., and Gool, L.V. (2016). Dynamic filter networks. Adv. Neural Inf. Process. Syst. *29*, 667–675.
14. Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE), pp. 4690–4699.
15. Zhang, N., Donahue, J., Girshick, R., and Darrell, T. (2014). Part-based r-cnns for fine-grained category detection. In European conference on computer vision (Springer), pp. 834–849.
16. Wei, X.-S., Xie, C.-W., Wu, J., and Shen, C. (2018). Mask-cnn: localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recognition *76*, 704–714.
17. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., and Wang, L. (2018). Learning to navigate for fine-grained classification. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 420–435.
18. Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (IEEE), pp. 7794–7803.
19. Cao, Y., Xu, J., Lin, S., Wei, F., and Hu, H. (2019). Gcnet: non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (IEEE).
20. Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (IEEE), pp. 7132–7141.
21. Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In European conference on computer vision (Springer), pp. 499–515.
22. Nech, A., and Kemelmacher-Shlizerman, I. (2017). Level playing field for million scale face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 7044–7053.
23. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. (2017). Sphereface: deep hypersphere embedding for face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (IEEE), pp. 212–220.
24. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. (2018). Cosface: large margin cosine loss for deep face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (IEEE), pp. 5265–5274.
25. He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (IEEE), pp. 1026–1034.
26. Kumar, A., Fulham, M., Feng, D., and Kim, J. (2019). Co-learning feature fusion maps from pet-ct images of lung cancer. IEEE Trans. Med. Imaging *39*, 204–217.
27. Zhu, Y., Liu, C., and Jiang, S. (2020). Multi-attention Meta Learning for Few-Shot Fine-Grained Image Recognition (IJCAI), pp. 1090–1096.
28. Woo, S., Park, J., Lee, J.-Y., and Kweon, I.S. (2018). Cbam: convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pp. 3–19.
29. Gao, Z., Xie, J., Wang, Q., and Li, P. (2019). Global second-order pooling convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE), pp. 3024–3033.

30. Qin, Z., Zhang, P., Wu, F., and Li, X. (2021). Fcanet: frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (IEEE), pp. 783–792.

31. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., and Huang, F. (2020). Curricularface: adaptive curriculum learning loss for deep face recognition. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition (IEEE), pp. 5901–5910.

32. Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (IEEE), pp. 2980–2988.

33. Ronneberger, O., Fischer, P., Brox, T., and U-net. (2015). Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (Springer), pp. 234–241.

34. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (IEEE), pp. 770–778.

35. Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (IEEE), pp. 1492–1500.

36. Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A.A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence (AAAI press).

37. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (IEEE), pp. 4700–4708.

38. Tan, M., and Le, Q. (2019). Efficientnet: rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (PMLR), pp. 6105–6114.

39. Ajaz, A., and Kathirvelu, D. (2013). Dental biometrics: computer aided human identification system using the dental panoramic radiographs. In 2013 international conference on communication and signal processing (IEEE), pp. 717–721.

40. Oktay, A.B. (2018). Human identification with dental panoramic radiographic images. IET Biometrics *7*, 349–355.