KINSHIP REPRESENTATION LEARNING WITH FACE COMPONENTIAL RELATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Kinship recognition aims to determine whether the subjects in two facial images are kin or non-kin, which is an emerging and challenging problem. However, most previous methods focus on heuristic designs without considering the spatial correlation between face images. In this paper, we aim to learn discriminative kinship representations embedded with the relation information between face components (e.g., eyes, nose, etc.). To achieve this goal, we propose the ***Fa**ce **Co**mponential **R**elation **Net**work (**FaCoRNet**)*, which learns the relationship between face components among images with a cross-attention mechanism, which automatically learns the important facial regions for kinship recognition. Moreover, we propose Relation-Guided Contrastive Learning, which adapts the loss function by the guidance from cross-attention to learn more discriminative feature representations. The proposed FaCoRNet outperforms previous state-of-the-art methods by large margins for the largest public kinship recognition FIW benchmark. The code will be publicly released upon acceptance.

1 INTRODUCTION

In recent years, *kinship recognition*, which aims to determine whether a given pair of face images have a kinship relation, has attracted public attention. Kinship recognition is inspired by the biological discovery (Dal Martello & Maloney (2010)) that the appearance of a human face implies clues about kinship-related information. It can be widely used in various scenarios including missing child search (Lu et al. (2013)), automatic album organization (Zhou et al. (2012)), child adoption (Yan et al. (2014)), and social media applications (Dehghan et al. (2014)). Facial kinship recognition includes both face representation learning and face similarity matching, where the former aims to learn discriminative features for input facial images, and the latter is to design models to predict the kin/non-kin relationship between images in a pair. The main challenges of kinship are mixed variations due to uncontrolled environment, such as the large gap in age, expression, pose, illumination, etc. Under these variations, it is challenging to learn representations that can help discover genetic relationships between two samples from facial appearance and identify hidden similarities inherited from genetic connections between different identities.

To deal with these challenges, several traditional approaches incorporate hand-crafted features (Lu et al. (2013)) with metric learning (Fan et al. (2020)) to learn discriminative features. Motivated by the success of deep learning, various methods improve kinship recognition by exploiting powerful deep feature representations. Zhang et al. (2015) first adopt a Convolutional Neural Networks (CNN) model to extract discriminative features, outperforming previous hand-crafted ones. For the extension, several CNN-based approaches (Luo et al. (2020); Dahan & Keller (2020); Yu et al. (2021)) focus on designing fusion mechanisms to integrate the features among an image pair. Recently, the supervised contrastive approach (Zhang et al. (2021)) learns discriminative features by contrastive loss, which achieves state-of-the-art performance in kinship recognition. However, the existing approaches have several issues. First, most methods directly exploit feature vector representations, ignoring spatial correlation within face images. Moreover, nearly all of the approaches rely on heuristic designs. For example, the feature fusion approaches (Zhang et al. (2015); Yu et al. (2020)) utilize several arithmetic combinations or feature concatenation to fuse the feature pair for kinship recognition. Despite state-of-the-art performance from Zhang et al. (2021), the results are sensitive in the hyperparameter setting of the contrastive loss.

To address the above issues, let us first think again: *How do humans recognize kinship relationships?* To recognize accurately, humans usually first compare several biological ***face components*** of two people, such as eye color, nose size, cheekbone shape, etc., and then analyze the ***relation*** between these comparisons. For example, if the noses (orange circle) in the image pair appear similarly, then there is a higher chance that this is a *kin* pair, as shown in Fig. 1 (left). Therefore, we adopt this idea, focusing on how to exploit these ***face components*** to learn the ***relation*** between images in a pair, where clues from *face components* can infer the genetic relationships between them. In this work, we aim to learn discriminative feature representations embedded with face component information, without a strong reliance on heuristic designs, as shown on the right-hand side of Fig. 1.

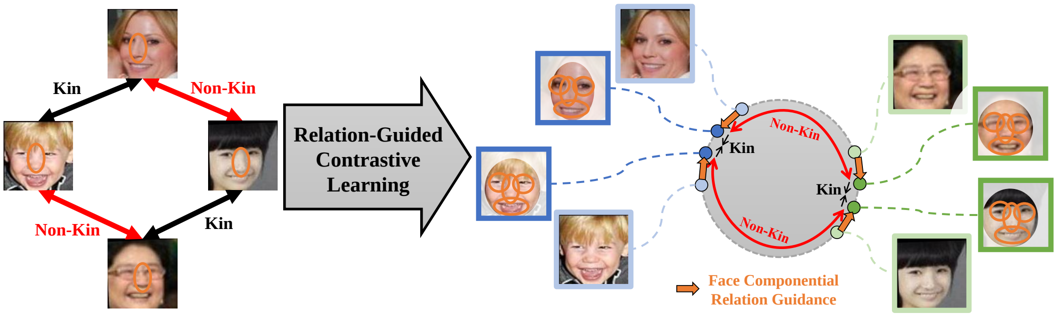
Kin
Non-Kin
Non-Kin
Kin
Relation-Guided Contrastive Learning
Non-Kin
Kin
Kin
Non-Kin
Face Componential Relation Guidance

Figure 1: The overall idea of this work. The left figure is an example of face components where noses in kinship image pairs are very similar and vice versa. The figure on the right shows that our method uses face components as clues and guides the training with the relation of facial image pairs, where the relation estimation for face components (darker blue and darker green colors) is more accurate than using face features alone (lighter blue and lighter green colors).

To achieve the abovementioned goal, we first propose the *Face Componential Relation (FaCoR)* module to learn the relation between images in a pair with the consideration of face components. The feature representations are then enhanced with the cross-relation between face components (e.g., eyes, nose, mouth, etc.) which are critical to kinship recognition. Moreover, we propose the novel *Relation-Guided Contrastive Loss* based on cross-attention estimation instead of heuristic tuning (Zhang et al. (2021)). The attention map can control the degree of penalty in the loss function, which can let the feature representation of kin relation get closer in the feature space. In other words, it penalizes the hard samples to learn more discriminative features for kinship recognition. The whole architecture is named ***F**ace **Co**mponential **R**elation **Net**work (**FaCoRNet**)*. The experimental results show that our FaCoRNet achieves SOTA performance on the largest public kinship recognition benchmark, FIW (Robinson et al. (2022)), our work outperforms the previous best method by 2.7% (79.3% → 82.0%) in the standard protocol and 3.4% (79.2% → 82.6%) in the practical protocol.

Our contributions are summarized as follows:

- We propose a novel ***F**ace **Co**mponential **R**elation **Net**work (**FaCoRNet**)* that learns relevance from the face components of image pairs with the cross-attention mechanism, and adaptively learns important face components for kinship recognition.
- We propose a novel *Relation-Guided Contrastive Loss* that embeds cross-relation estimates to guide the contrastive loss without heuristic tuning, which controls how hard samples are penalized during the training phase.
- The proposed ***FaCoRNet*** model outperforms previous SOTA methods by large margins in the largest kinship recognition benchmark.

2 Related Work

In the past few years, several kinship recognition approaches have been proposed (Lu et al. (2013); Zhang et al. (2015); Fan et al. (2020); Dahan & Keller (2020); Song & Yan (2020); Hörmann et al. (2020); Luo et al. (2020); Shadrikov (2020); Lin et al. (2021); Yu et al. (2021); Zhang et al. (2021)), most of them focus on extracting discriminative feature for each facial image. Traditional approaches include designing hand-crafted feature extractors (e.g., PCA (Abdi & Williams (2010)), SIFT (Somanath & Kambhamettu (2012)), SFRD (Cui et al. (2013)), etc.), and metric learning (Ding & Tao (2017); Dibeklioglu (2017)) for solving similarity metrics in kinship recognition. Recently, deep learning approaches make significant advances and can be divided into two main categories: *feature fusion* and *deep metric learning*.

Feature fusion: (Zhang et al. (2015)) utilizes the multiple face regions into the model input to learn richer facial features for kinship recognition. The multi-task deep learning-based approach (Dahan & Keller (2020)) uses all 7 kinship sub-classes to jointly train with the kinship labels for kin recognition. Ustc-nelslip (Yu et al. (2020)) adopts a shared-weights multi-model to extract features and designs three different math operations to fuse feature pairs. All the features are then concatenated directly, followed by a fully-connected layer.

Deep metric learning: Duan et al. (2017) proposes coarse-to-fine transfer to capture kinship-specific features from faces using supervised coarse pre-training and domain-specific retraining paradigms. The contrastive learning approach (Zhang et al. (2021)) utilizes supervised contrastive loss with the ArcFace pre-trained model (Deng et al. (2019)) and two MLP layers to learn more robust features in the training stage. For the evaluation, it removes the MLP layers and extracts the middle-layer backbone features to evaluate the cosine similarity to determine the kinship relation between images in a pair, currently achieving state-of-the-art performance for kinship recognition.

The main issues of the above methods are that most methods rely on heuristic designs, and directly exploit feature vector representations, ignoring spatial correlation within face images. Different from the above approaches, our proposed FaCoRNet considers how to use face components to learn the correlation between image pairs, in particular, which facial parts are important for kinship recognition. Moreover, our approach incorporates the correlation information from face components to adapt contrastive learning automatically, without a strong reliance on heuristic designs.

3 Proposed Methods

In this work, we propose the ***Fa**ce **Co**mponential **R**elation **Net**work (**FaCoRNet**)*, which considers the *face components* and learn the cross-*relation* between face images in a pair to benefit kinship recognition. FaCoRNet consists of a shared-weights backbone that extracts features as the inputs to the *Face Componential Relation (FaCoR)* module. FaCoR is an attention-based module that calculates the cross-relation among a face image pair and enhances feature representations to fully exploit the symmetry of face components in the image pair (Sec. 3.1). In the FaCoR module, cross-layer features are interacted and fused in the channel dimension by the *Channel Interaction (CI)* block (Sec. 3.1.1). Finally, the proposed *Relation-Guided Contrastive Loss* utilizes the computed cross-relation to guide the contrastive loss, facilitating learning of more discriminative representations for kinship recognition (Sec. 3.2). The overall framework is illustrated in Fig. 2.

3.1 Face Componential Relation (FaCoR)

One core question for kinship recognition is: *How to properly extract and calculate the relation between face components among a face image pair?* However, most existing methods are not designed for the face components of kinship recognition. To solve this, we propose the Face Componential Relation (FaCoR) module, which can embed the relation information between face components into kinship feature representations, as the core component of our FaCoRNet (Fig. 2).

We denote the input face image pair as $(\mathbf{I}^a, \mathbf{I}^b) \in \mathbb{R}^{h \times w \times 3}$, the extracted feature maps from the backbone's middle-layer as $(\mathbf{X}^a, \mathbf{X}^b) \in \mathbb{R}^{H \times W \times C}$, and the high-level features from the backbone's final layer as $(\mathbf{r}^a, \mathbf{r}^b) \in \mathbb{R}^C$, where H, W, and C represent the height, width, and the channel number of feature maps, respectively. The proposed FaCoR module mainly serves two purposes: 1) To adaptively learn the correlation between face image pairs, and 2) to learn the dependencies in face components between image pairs. These two directions help to learn which facial parts are important for kinship recognition. More specifically, We first extract features $(\mathbf{X}^a, \mathbf{X}^b)$ from the shared-weights backbone and then use 1×1 convolution Conv to extract two intermediate flattened feature vectors $(\mathbf{F}^a, \mathbf{F}^b) = (\mathrm{Conv}_{1\times1}(\mathbf{X}^a), \mathrm{Conv}_{1\times1}(\mathbf{X}^b)) \in \mathbb{R}^{H \times W \times C}$. Then, we find wide-range dependencies between the flattened feature vector pair $(\mathbf{F}^a, \mathbf{F}^b)$ and estimate the cross-

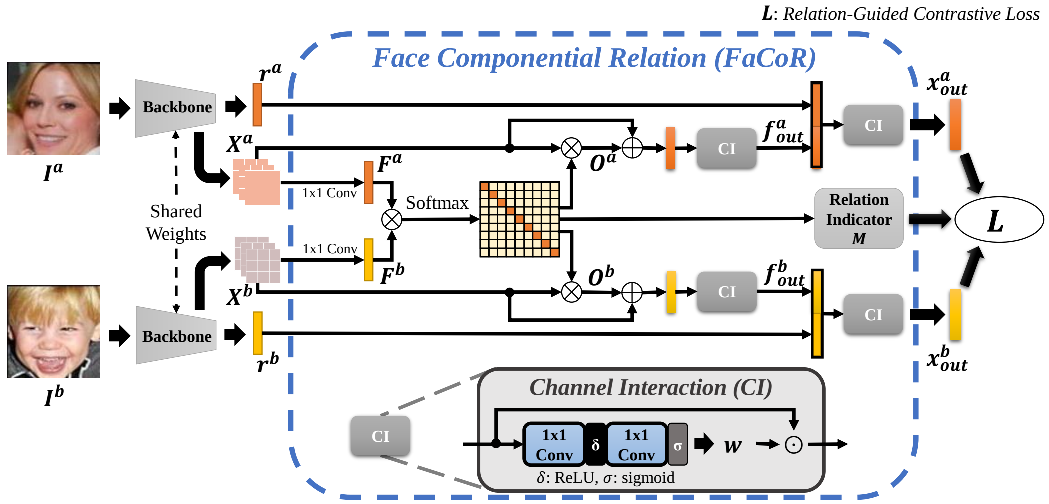

Figure 2: An overview of the proposed *Face Componential Relation Network (FaCoRNet)* consisting of a backbone and the Face Componential Relation (FaCoR) module (Sec. 3.1) including the Channel Interaction (CI) block (Sec. 3.1.1), trained with the Relation-Guided Contrastive Loss L (Sec. 3.2).

attention map $\boldsymbol{\beta}$ as:

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N}\exp(s_{ij})}\ ,\ s_{ij} = (\mathbf{F}_i^a)^T\mathbf{F}_j^b \tag{1}$$

where $\beta_{j,i}$ estimates model attention in the i-th location of the j-th region.

We then multiply each output of the attention map $\boldsymbol{\beta}$ with the feature map $(\mathbf{X}^a, \mathbf{X}^b)$ to obtain the cross-attention features $(\mathbf{O}^a, \mathbf{O}^b) \in \mathbb{R}^{C\times HW}$ given by:

$$\left(\mathbf{O}_j^a,\ \mathbf{O}_j^b\right) = \left(\sum_{i=1}^{N}\beta_{j,i}\mathbf{X}_i^a,\ \sum_{i=1}^{N}\beta_{j,i}\mathbf{X}_i^b\right) \tag{2}$$

All the operations are differentiable since they are purely linear and properly reshaped.

We further adopt a scaled residual connection to the cross-attention features $\left(\mathbf{O}^a, \mathbf{O}^b\right)$ with a learnable scaling parameter γ. Then a Channel Interaction (CI) block, which utilizes the attention mechanism in full channel dimension, is adopted to generate output pair $\left(\mathbf{f}^{\mathbf{a}}_{\mathbf{out}}, \mathbf{f}^{\mathbf{b}}_{\mathbf{out}}\right)$ as:

$$\left(\mathbf{f}^{\mathbf{a}}_{\mathbf{out}}, \mathbf{f}^{\mathbf{b}}_{\mathbf{out}}\right) = \left(\mathbf{CI}\left(\gamma\mathbf{O}^a + \mathbf{X}^a\right), \mathbf{CI}\left(\gamma\mathbf{O}^b + \mathbf{X}^b\right)\right) \tag{3}$$

where **CI** is the Channel Interaction (CI) block and will be elaborated later in Sec. 3.1.1.

Finally, we utilize another CI block to fuse the information from the high-level features $(\mathbf{r}^a, \mathbf{r}^b)$ and the cross-attention features $\left(\mathbf{f}^{\mathbf{a}}_{\mathbf{out}}, \mathbf{f}^{\mathbf{b}}_{\mathbf{out}}\right)$, generating the final outputs $(\mathbf{x}^{\mathbf{a}}_{\mathbf{out}}, \mathbf{x}^{\mathbf{b}}_{\mathbf{out}})$ as:

$$\left(\mathbf{x}^{\mathbf{a}}_{\mathbf{out}}, \mathbf{x}^{\mathbf{b}}_{\mathbf{out}}\right) = \left(\mathbf{CI}\left(\mathbf{f}^{\mathbf{a}}_{\mathbf{out}} \mid\mid \mathbf{r}^a\right), \mathbf{CI}\left(\mathbf{f}^{\mathbf{b}}_{\mathbf{out}} \mid\mid \mathbf{r}^b\right)\right) \tag{4}$$

where the operation $||$ denotes the concatenation of two feature maps in the channel dimension.

3.1.1 CHANNEL INTERACTION (CI)

To effectively fuse the information from different features, we propose a Channel Interaction (CI) block that utilizes full channel attention as shown in the gray block in Fig. 2. CI computes the interaction weights $\mathbf{w}$ via two sets of 1×1 convolution, a sigmoid, and a ReLU activation function as:

$$\mathbf{w} = \sigma \left(\mathrm{Conv}_{1\times1} \left(\delta \left(\mathrm{Conv}_{1\times1}(\hat{\mathbf{x}}) \right) \right) \right) \tag{5}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ is a 1×1 convolution operation, σ is the sigmoid operation, and δ is the ReLU operation. $\hat{\mathbf{x}}$ denotes the input of the CI block.

In our FaCoR module, we have two types of CI blocks: *non-group* CI (used in Eq. 3) and *group* CI (used in Eq. 4). The non-group CI block first applies the global average pooling (GAP) operation to obtain the feature vectors as input, and then directly multiplies the interaction weights $\mathbf{w}$ with the input features element-wisely. The group CI block is used to effectively fuse cross-attention features $\left(\mathbf{f^a_{out}}, \mathbf{f^b_{out}}\right)$ with the high-level features $(\mathbf{r}^a, \mathbf{r}^b)$ in the channel dimension to perform group channel interaction. More specifically, we divide the interaction weights $\mathbf{w}$ into two groups $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2]$ corresponding to two different groups of features (i.e., the cross-attention features $\left(\mathbf{f^a_{out}}, \mathbf{f^b_{out}}\right)$ and the high-level features $(\mathbf{r}^a, \mathbf{r}^b)$, respectively), and we conduct an element-wise product separately to the two groups of features with the corresponding weights. Finally, we sum up the two groups of features to become the final output $\mathbf{x^a_{out}} = \mathbf{w}_1 \mathbf{f_{out}} + \mathbf{w}_2 \mathbf{r}$.

3.2 Relation-Guided Contrastive Learning (Rel-Guide)

Contrastive learning (Chen et al. (2020); Khosla et al. (2020)) is known as an effective representation learning approach. It allows the model to learn the discriminative features from data similarities and dissimilarities, even without labels. The supervised contrastive (Zhang et al. (2021)) approach learns more robust features in kinship recognition, which achieve state-of-the-art performance. The main idea of contrastive learning is to learn the discriminative feature, where feature representations of kin relations in feature space would be nearby. Otherwise, the feature representations of non-kin relation in feature space are far apart. For the standard contrastive learning, given N positive samples x_i, y_i, the contrastive loss L is formulated as:

$$L = \frac{1}{2N} \left(\sum_{i=1}^{N} L_c(x_i, y_j) + L_c(y_i, x_i) \right) \tag{6}$$

and $L_c(x_i, y_i)$ is defined as:

$$L_c(x_i, y_i) = -log \frac{e^{sim(x_i, y_i)/\tau}}{\sum_{j=1}^{N} \left(e^{sim(x_i, x_j)/\tau} + e^{sim(x_i, y_j)/\tau} \right)} \tag{7}$$

where $sim(x, y)$ is the cosine similarity operation between x and y, and the negative samples are generated by incorporating positive from different kinship categories.

However, the kinship recognition performance of contrastive learning is sensitive to hyper-parameter τ (Zhang et al. (2021)), which controls the degree of penalty for hard samples. The smaller τ penalizes hard samples to a greater degree and vice versa. To solve this problem, we propose the *Relation-Guided Contrastive Loss* by a relation indicator as shown in Fig. 2. The relation indicator guides the contrastive loss from the cross-attention estimation instead of heuristic tuning in τ.

The main idea is that a smaller value from the cross-attention map needs a greater degree of penalty for hard samples. In other words, the small correlation between image pairs in kin relation needs a greater degree of penalty. This idea is also similar to updating the network with a large gradient to improve kinship recognition performance, and vice versa. Therefore, we extract the cross-attention map $\boldsymbol{\beta}$ in Eq. 1, which corresponds to the face component correlation between image pairs. Then, we utilize the relation indication function $\mathbf{M}$ to estimate the similarity value ψ to replace the fixed value of τ in Eq. 8 as:

$$L_c(x_i, y_i) = -log \frac{e^{sim(x_i, y_y)/\psi}}{\sum_{j=1}^{N} \left(e^{sim(x_i, x_j)/\psi} + e^{sim(x_i, y_j)/\psi} \right)}, \quad \psi = \mathbf{M}(\boldsymbol{\beta})/s \tag{8}$$

where s is the scale value and we adopt the global sum pooling operation as the relation indicator $\mathbf{M}$. In our FaCoRNet, the feature pair (x, y) in loss function use the output feature pairs $(\mathbf{x}^{\mathbf{a}}_{\mathbf{out}}, \mathbf{x}^{\mathbf{b}}_{\mathbf{out}})$ from Face Componential Relation to calculate loss for updating the model.

In the inference, we follow (Zhang et al. (2021)) for the training and inference process. We extract the final outputs $(\mathbf{x}^{\mathbf{a}}_{\mathbf{out}}, \mathbf{x}^{\mathbf{b}}_{\mathbf{out}})$ from Eq. 4 to calculate the cosine similarity, and a threshold is used to predict whether there is a kinship relation between them.

4 Experiments

4.1 Dataset and Evaluation: Kinship Recognition

The compared methods are all trained and tested on a publicly available kinship recognition dataset **Families in the Wild (FIW) (Robinson et al. (2022))**. The FIW dataset is the largest kinship recognition dataset which includes 1000 different and disjoint family trees, around 12000 family photos, and 11 kin relationship types. All face images are cropped to the size of 112 × 112 with face detection and alignment in training and testing by MTCNN (Zhang et al. (2016)). The 11 kin relationship types include: a) *Siblings*: Brother-Brother (BB), Sister-Sister (SS), and Sister-Brother (SIBS); b) *Parent-Child*: Father-Daughter (FD), Mother-Daughter (MD), Father-Son (FS), and Mother-Son (MS); c) *Grandparent-Grandchild*: GFGD, GFGS, GMGD, and GMGS, with the same naming convention as above. In this work, we mainly focus on the first 7 kinship relationships since the Grandparent-grandchild categories contain much smaller data by an order of magnitude, as shown in Fig. 3. For evaluation, we adopt cosine similarity and thresholding to calculate accuracy as following the FIW benchmark (Robinson et al. (2022)).

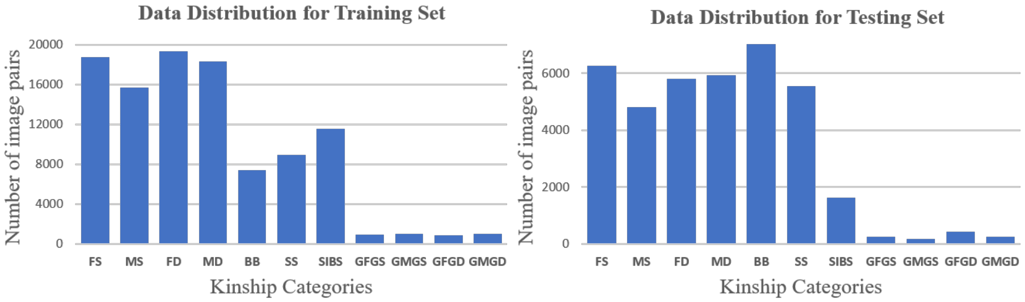

Figure 3: Illustration of the data distribution on the FIW dataset. The x-axis is the 11 kinship categories and the y-axis is the number of image pairs, respectively. The left and right figures represent the data distribution for the training and testing set, respectively.

4.2 Implementation details

For experiments, we select 103784 positive and negative image pairs overall without non-aligned images for training phase and follow the evaluation protocols as detailed in (Robinson et al. (2017)) by applying the restricted protocol where the identities of the subjects in the dataset are unknown, and we are given predefined pairs of training images, per kinship class. We compare our FaCoR-Net against several existing methods by using ArcFace (Deng et al. (2019)) as the pre-trained backbone for a fair comparison. To demonstrate the advanced face feature representation for kinship recognition, we use the SOTA face recognition model, AdaFace (Kim et al. (2022)), as the feature extraction network to compare with SOTA kinship recognition methods. Since the naive pre-trained weights of Adaface are not suitable for the kinship method (more details in Sec. 4.3.1), we modified the initialization model parameters as a normal distribution, with lower and upper bound to [-0.05, 0.05] and utilize the L2-norm feature normalization. For the training scheme, We use SGD as the optimizer with a constant learning rate of 1e-4 and a momentum of 0.9. The batch size is set to 50, and the total number of iterations to 4000 for 50 epochs.

4.3 Experimental Results

4.3.1 Comparison to SOTA Methods

We first evaluate kinship recognition performance with the FIW dataset under the standard protocol, and compare our proposed FaCoRNet with several state-of-the-art methods including stefhoer (Hörmann et al. (2020)), DeepBlueAI (Luo et al. (2020)), Vuvko (Shadrikov (2020), Ustc-nelslip (Yu et al. (2020)), and Contrastive (Zhang et al. (2021)). Table 1 compares the kinship recognition accuracy by using two different pre-trained models (i.e., ArcFace and AdaFace) by various methods. The result shows that the kinship recognition average accuracy from our proposed method is significantly higher than those achieved by the other methods. For the standard comparison which adopts Arcface (Deng et al. (2019)) as the pre-trained model, our FaCoRNet outperforms previous leading methods Ustc-nelslip, Vuvko, and Contrastive by 4.6 percent (0.760 → 0.806), 2.6 percent (0.780 → 0.806), and 1.3 percent (0.793 → 0.806), respectively, as shown in Table 1(a). Then a question arises: *Do advanced face recognition models benefit kinship recognition?* To answer this, we adopt Contrastive (Zhang et al. (2021)) as the strong baseline and exploit AdaFace (Kim et al. (2022)) as pre-trained for better general initial face representation. However, naively replacing the pre-trained model with Adaface is not suitable for kinship recognition, as the average accuracy decrease significantly (0.793 → 0.728). We then modify the training scheme as stated in Sec. 4.2 and show that advanced face recognition models can improve the kinship recognition task (0.793 → 0.802). Finally, by integrating the modified AdaFace backbone with our proposed FaCoRNet, the result is further boosted by 1.8 percent (0.802 → 0.820), achieving the SOTA performance (Table 1(b)). To summarize, by integrating the advanced face recognition model with FaCoRNet and our proposed training scheme, our result significantly outperforms the previous SOTA method by 2.7 percent (0.793 → 0.820), achieving a new SOTA result.

Table 1: The state-of-the-art performance comparison of kinship recognition on FIW dataset by two pre-trained backbones: (a) ArcFace (Deng et al. (2019)) and (b) AdaFace (Kim et al. (2022)). The best accuracy in each column is highlighted in bold. †The results are from Robinson et al. (2022).

Method	BB	SS	SIBS	FD	MD	FS	MS	AVG.
(a) Pre-trained model: ArcFace (Deng et al. (2019))								
Stefhoer† (Hörmann et al. (2020))	0.660	0.650	0.760	0.770	0.770	0.800	0.780	0.740
DeepBlueAI† (Luo et al. (2020))	0.770	0.770	0.750	0.740	0.750	0.810	0.740	0.760
Ustc-nelslip† (Yu et al. (2020))	0.750	0.740	0.720	0.760	0.750	0.820	0.750	0.760
Vuvko† (Shadrikov (2020))	0.800	0.800	0.770	0.750	0.780	0.810	0.740	0.780
Contrastive (Zhang et al. (2021))	0.803	0.829	0.794	0.753	0.803	0.823	0.751	0.793
FaCoRNet (Ours)	**0.824**	0.827	0.804	0.763	**0.806**	0.824	0.779	0.803
FaCoRNet + Rel-Guide (Ours)	0.820	**0.833**	**0.810**	**0.773**	0.804	**0.826**	**0.788**	**0.806**
(b) Pre-trained model: AdaFace (Kim et al. (2022))								
Contrastive (Zhang et al. (2021))	0.630	0.776	0.731	0.663	0.687	0.736	0.687	0.728
Contrastive (Zhang et al. (2021)) (modified)	0.821	0.831	0.798	0.766	0.806	0.828	0.767	0.802
FaCoRNet (Ours)	**0.832**	**0.840**	0.821	0.790	**0.822**	**0.848**	**0.802**	0.819
FaCoRNet + Rel-Guide (Ours)	**0.832**	0.836	**0.824**	**0.795**	0.818	**0.848**	**0.802**	**0.820**

4.3.2 Practical Kinship Recognition Protocol

With the improvement of hardware, the photos captured by cameras or smartphones have better quality, so it is worth investigating the impact of using higher-quality face images in practical applications. We conduct an experiment for practical kinship recognition as shown in Table 2 (b). More specifically, we propose a quality-filtered protocol, where we select high-quality training and testing face images with SER-FIQ quality scores (Terhorst et al. (2020)) larger than 0.5. The results demonstrate that the average accuracy of FaCoRNet are significantly higher than the baseline (i.e., Contrastive). This trend is similar to the standard protocol as shown in Table 2 (a), but the improvement from our method over the baseline is even more obvious (0.792 → 0.826).

Intuitively, using high-quality face images as training and testing data would improve overall accuracy. However, the accuracy of Contrastive (Zhang et al. (2021)) does not improve on high-quality face images, which also confirms that our FaCoRNet can learn the correlation between image pairs and fuse them more effectively, that is, capture the face components from eye, nose, and mouth. Besides, we further analyze the recognition results of different kinships and found that the accuracy of the same gender (i.e., BB, SS, MD, FS) was significantly higher. Among them, the result of FaCoRNet + Rel-Guide in MD case has a significant improvement of 2.4 percent (0.818 → 0.842) from the standard to the quality-filtered protocol, showing that the MD cases include a large amount of low-quality face images in the standard protocol. On the other hand, MD has slightly lower recognition accuracy than FS, and we conjecture that it is due to the challenging MD cases caused by makeup and coverings as shown in Fig. 4. Moreover, the accuracy of the SIBS case decreases after selecting high-quality face images. The main reason is that SIBS has less data than other kinship categories as illustrated in Fig. 3. Finally, the results also demonstrate that our FaCoRNet outperforms the previous SOTA method by a large margin in all kin categories.

Table 2: Performance comparison of kinship on FIW dataset in two quality-filtered protocols: (a) standard protocol: use all image pairs without filtering; (b) quality-filtered protocol: select the image pairs with the pair quality scores larger than 0.5, which is more practical in real-world scenarios.

Method	BB	SS	SIBS	FD	MD	FS	MS	AVG.
(a) Standard Protocol								
Contrastive (Zhang et al. (2021))	0.803	0.829	0.794	0.753	0.803	0.823	0.751	0.793
FaCoRNet (AdaFace) (Ours)	**0.832**	**0.840**	0.821	0.790	**0.822**	**0.848**	**0.802**	0.819
FaCoRNet + Rel-Guide (AdaFace) (Ours)	**0.832**	0.836	**0.824**	**0.795**	0.818	**0.848**	**0.802**	**0.820**
(b) Quality-Filtered Protocol (Quality Score > 0.5)								
Contrastive (Zhang et al. (2021))	0.800	0.817	0.772	0.739	0.784	0.836	0.786	0.792
FaCoRNet (AdaFace) (Ours)	0.834	**0.842**	0.780	0.770	0.839	**0.865**	0.804	0.823
FaCoRNet + Rel-Guide (AdaFace) (Ours)	**0.836**	0.838	**0.784**	**0.784**	**0.842**	0.862	**0.815**	**0.826**

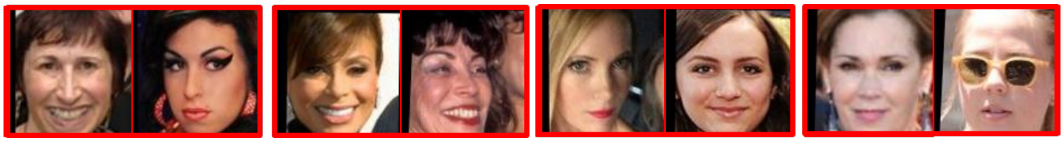

Figure 4: Illustration of the hard samples in the Mother-Daughter (MD) case. This figure shows that the face has makeup, glasses, etc., which makes it challenging to identify the kin relation.

4.3.3 Ablation Studies

Component Analysis: In this section, we conduct an ablation study to analyze the proposed design for comparisons against various component modules. Our proposed FaCoRNet utilize the Face Componential Relation (FaCoR) to extract the face components of both images in a pair by assigning the important facial regions for kinship recognition. Table 3 reports the contributions of individual modules of FaCoRNet. The first row shows the naive approach of contrastive learning with 2 fully-connected layers from Zhang et al. (2021) only. The second raw shows the improvement of utilizing (Zhang et al. (2021)) with our proposed FaCoR module. The result demonstrates that our FaCoR module has a significant improvement, which reveals that FaCoR can extract the face component information by learning the correlation between image pairs. In the third row, the channel Channel Interaction (CI) is included for FaCoR and boost the average accuracy, which shows it works better for fusing the cross-attention features with the high-level features from the backbone output. The last row shows that the relation guidance in contrastive loss can further improve kinship performance.

Table 3: Component analysis of FaCoRNet (pre-trained on ArcFace) on the FIW dataset.

Contrastive	Face Componential Relation	Channel Interaction	Relation Guidance	AVG.
✓				0.793
✓	✓			0.795
✓	✓	✓		0.803
✓	✓	✓	✓	**0.806**

Face Quality Analysis: We do a comparison experiment with the case of the various face quality. We utilize SER-FIQ (Terhorst et al. (2020)) to compute the face quality scores of all images and adopt the lower score in a pair as the face-pair quality score. We divide the face-pair quality scores into 5 groups as shown in Table 4. The results show that in low-quality cases (i.e., the quality scores are smaller than 0.4), the overall recognition accuracy is lower than in high-quality cases. The problem is more severe in extremely low-quality cases (i.e., 0.2). Finally, the results also demonstrate that our FaCoRNet outperforms the SOTA method under all quality cases.

Table 4: Performance comparison of kinship on FIW dataset under various groups of pair quality scores. The table represents the pair quality score in groups from small to large.

Face-Pair Quality Score	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1	AVG.
Contrastive (Zhang et al. (2021))	0.749	0.782	0.813	0.803	0.793	0.792
FaCoRNet (AdaFace) (Ours)	**0.794**	**0.820**	**0.843**	**0.821**	**0.824**	**0.820**

Limitations: We enumerate some of the failure cases in Fig. 5. Among these hard samples, some face pairs are captured in different scenarios and ages, resulting in large variants in illumination, expression, and pose; except for the low-quality face image case as mentioned above, some of them are obscured and covered (i.e., bread and wearing glasses), making the kinship recognition challenging. Besides, in the case of low-quality face images, as long as one image in the pair is of poor quality, the recognition result will be seriously affected. Finally, extreme poses also cause difficulty for kinship recognition due to less face component information.

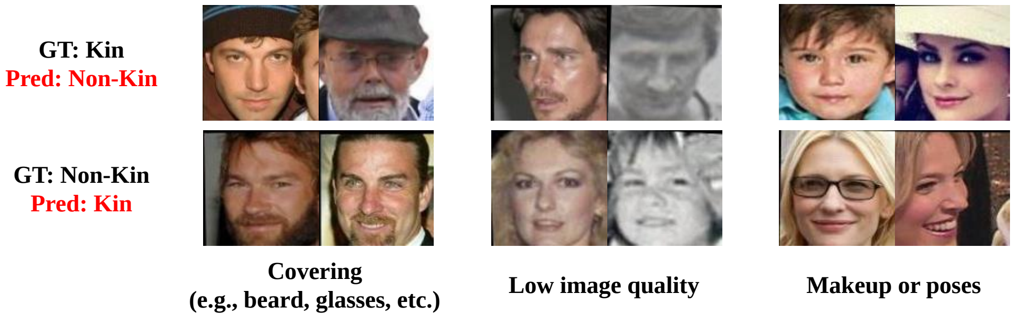

Figure 5: Illustration of failure cases on FIW dataset include: The left side shows the cases where the face is covered by a beard, glasses, etc; The middle shows low-quality images of a face; The right-hand side shows the face with makeup and different poses.

5 Conclusion and Future Work

In this paper, we propose a novel ***Face Componential Relation Network (FaCoRNet)*** for kinship recognition. FaCoRNet is an attention-based model designed for learning correlation between image pairs in terms of face components. To better address large variations in facial appearance, FaCoRNet utilizes the *Face Componential Relation (FaCoR)* module to achieve not only adaptive learning correlation between image pairs but also learning important face components for kinship recognition. In addition, we embed the cross-attention estimation as a relation indicator to guide the regular contrastive loss without the need for heuristic tuning, which can control the degree of penalty for hard samples in the training phase. Experimental results show that our method achieves SOTA performance on the largest public kinship recognition FIW benchmark. Moreover, in terms of practical kinship recognition protocol, FaCoRNet also outperforms previous SOTA methods by large margins. We believe that FaCoRNet is a potential kinship recognition method that can be served as a strong baseline for further advancing facial relation learning approaches in kinship recognition. For future work, we plan to incorporate face quality scores into the training process, aiming to mitigate the issues from low-quality face images. We would also like to incorporate multi-modal information (e.g., text, metadata) to compensate for the vision-based methods.

References

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Zhen Cui, Wen Li, Dong Xu, Shiguang Shan, and Xilin Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3554–3561, 2013.

Eran Dahan and Yosi Keller. A unified approach to kinship verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2851–2857, 2020.

Maria F Dal Martello and Laurence T Maloney. Lateralization of kin recognition signals in the human face. *Journal of vision*, 10(8):9–9, 2010.

Afshin Dehghan, Enrique G Ortiz, Ruben Villegas, and Mubarak Shah. Who do i look like? determining parent-offspring resemblance via gated autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1757–1764, 2014.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Hamdi Dibeklioglu. Visual transformation aided contrastive learning for video-based kinship verification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2459–2468, 2017.

Changxing Ding and Dacheng Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):1002–1014, 2017.

Qingyan Duan, Lei Zhang, and Wangmeng Zuo. From face recognition to kinship verification: An adaptation approach. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 1590–1598, 2017.

Bin Fan, Qingqun Kong, Baoqian Zhang, Hongmin Liu, Chunhong Pan, and Jiwen Lu. Efficient nearest neighbor search in high dimensional hamming space. *Pattern Recognition*, 99:107082, 2020.

Stefan Hörmann, Martin Knoche, and Gerhard Rigoll. A multi-task comparator framework for kinship verification. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 863–867. IEEE, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18750–18759, 2022.

Che-Hsien Lin, Hung-Chun Chen, Li-Chen Cheng, Shu-Chuan Hsu, Jun-Cheng Chen, and Chih-Yu Wang. Styledna: A high-fidelity age and gender aware kinship face synthesizer. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–8. IEEE, 2021.

Jiwen Lu, Xiuzhuang Zhou, Yap-Pen Tan, Yuanyuan Shang, and Jie Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):331–345, 2013.

Zhipeng Luo, Zhiguang Zhang, Zhenyu Xu, and Lixuan Che. Challenge report recognizing families in the wild data challenge. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 868–871. IEEE, 2020.

Joseph P Robinson, Ming Shao, Yue Wu, Hongfu Liu, Timothy Gillis, and Yun Fu. Visual kinship recognition of families in the wild (fiw). *IEEE Transactions on Pattern Analysis and Machine Intelligence Special Issue: The Computational Face*, 2017.

Joseph P Robinson, Ming Shao, and Yun Fu. Survey on the analysis and modeling of visual kinship: A decade in the making. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4432–4453, 2022.

Andrei Shadrikov. Achieving better kinship recognition through better baseline. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 872–876. IEEE, 2020.

Gowri Somanath and Chandra Kambhamettu. Can faces verify blood-relations? In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 105–112. IEEE, 2012.

Chaohui Song and Haibin Yan. Kinmix: A data augmentation approach for kinship verification. In *2020 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6. IEEE, 2020.

Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5651–5660, 2020.

Haibin Yan, Jiwen Lu, Weihong Deng, and Xiuzhuang Zhou. Discriminative multimetric learning for kinship verification. *IEEE Transactions on Information forensics and security*, 9(7):1169–1178, 2014.

Jun Yu, Mengyan Li, Xinlong Hao, and Guochen Xie. Deep fusion siamese network for automatic kinship verification. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 892–899. IEEE, 2020.

Jun Yu, Guochen Xie, Xinlong Hao, Zeyu Cui, Liwen Zhang, and Zhongpeng Cai. Deep kinship verification and retrieval based on fusion siamese neural network. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–8. IEEE, 2021.

Kaihao Zhang, Yongzhen Huang, Chunfeng Song, Hong Wu, Liang Wang, and Statistical Machine Intelligence. Kinship verification with deep convolutional neural networks. In *The British Machine Vision Conference*. British machine vision conference. BMVA Press, 2015.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

Ximiao Zhang, XU Min, Xiuzhuang Zhou, and Guodong Guo. Supervised contrastive learning for facial kinship recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 01–05. IEEE, 2021.

Xiuzhuang Zhou, Jiwen Lu, Junlin Hu, and Yuanyuan Shang. Gabor-based gradient orientation pyramid for kinship verification under uncontrolled environments. In *Proceedings of the 20th ACM international conference on Multimedia*, pp. 725–728, 2012.

6 Appendix

6.1 Supplementary Results

We have supplemented a full comparison of all 11 kinship categories in Table 5 and Table 6, including the addition of *Grandparent-Grandchild* categories: GFGD, GFGS, GMGD, and GMGS.

Table 5: The state-of-the-art performance comparison of kinship recognition on FIW dataset in all 11 kinship categories by two pre-trained backbones: (a) ArcFace (Deng et al. (2019)) and (b) AdaFace (Kim et al. (2022)). The best accuracy in each column is highlighted in bold. †The results are from Robinson et al. (2022).

Method	BB	SS	SIBS	FD	MD	FS	MS	GFGD	GMGD	GFGS	GMGS	AVG.
(a) Pre-trained model: ArcFace (Deng et al. (2019))												
Stefhoer† (Hörmann et al. (2020))	0.660	0.650	0.760	0.770	0.770	0.800	0.780	0.700	0.640	0.730	0.600	0.740
DeepBlueAI† (Luo et al. (2020))	0.770	0.770	0.750	0.740	0.750	0.810	0.740	0.720	0.670	0.730	0.680	0.760
Ustc-nelslip† (Yu et al. (2020))	0.750	0.740	0.720	0.760	0.750	0.820	0.750	0.790	0.760	0.690	0.670	0.760
Vuvko† (Shadrikov (2020))	0.800	0.800	0.770	0.751	0.780	0.810	0.740	0.780	0.760	0.690	0.690	0.780
Contrastive (Zhang et al. (2021))	0.803	0.829	0.794	0.753	0.803	0.823	0.751	0.754	0.740	0.702	0.592	0.793
FaCoRNet (Ours)	0.82	0.827	0.804	0.763	0.806	0.824	0.779	0.756	0.703	0.698	0.592	0.803
FaCoRNet + Rel-Guide (Ours)	0.820	0.833	0.810	0.773	0.804	0.826	0.788	0.774	0.706	0.702	0.587	**0.806**
(b) Pre-trained model: AdaFace (Kim et al. (2022))												
Contrastive (Zhang et al. (2021))	0.630	0.776	0.731	0.663	0.687	0.736	0.687	0.722	0.665	0.669	0.525	0.728
Contrastive (Zhang et al. (2021)) (modified)	0.821	0.831	0.798	0.766	0.806	0.828	0.767	0.756	0.725	0.669	0.626	0.802
FaCoRNet (Ours)	0.832	0.840	0.821	0.790	0.822	0.848	0.802	0.806	0.684	0.694	0.575	0.819
FaCoRNet + Rel-Guide (Ours)	0.832	0.836	0.824	0.795	0.818	0.848	0.802	0.799	0.684	0.690	0.575	**0.820**

Table 6: Performance comparison of kinship on FIW dataset in all 11 kinship categories for two quality-filtered protocols: (a) standard protocol: use all image pairs without filtering; (b) quality-filtered protocol: select the image pairs with the pair quality scores larger than 0.5, which is more practical in real-world scenarios.

Method	BB	SS	SIBS	FD	MD	FS	MS	GFGD	GMGD	GFGS	GMGS	AVG.
(a) Standard Protocol												
Contrastive (Zhang et al. (2021))	0.803	0.829	0.794	0.753	0.803	0.823	0.751	0.754	0.740	0.702	0.592	0.793
FaCoRNet (AdaFace) (Ours)	0.832	0.840	0.821	0.790	0.822	0.848	0.802	0.806	0.684	0.694	0.575	0.819
FaCoRNet + Rel-Guide (AdaFace) (Our)	0.832	0.836	0.824	0.795	0.818	0.848	0.802	0.799	0.684	0.690	0.575	**0.820**
(b) Quality-Filtered Protocol (Quality Score > 0.5)												
Contrastive (Zhang et al. (2021))	0.800	0.817	0.772	0.739	0.784	0.836	0.786	0.791	0.737	0.703	0.691	0.792
FaCoRNet (AdaFace) (Ours)	0.834	0.842	0.780	0.770	0.839	0.865	0.804	0.809	0.766	0.717	0.639	0.823
FaCoRNet + Rel-Guide (AdaFace) (Ours)	0.836	0.838	0.784	0.784	0.842	0.862	0.815	0.810	0.686	0.715	0.651	**0.826**

Since the Grandparent-Grandchild categories have only one-tenth of the data of the other categories, there is not enough data for model training and inference This is the potential reason why our FaCoRNet has sub-optimal performance in the Grandparent-Grandchild categories.