Kinship Representation Learning with Face Componential Relation

Wen-Tai Su*
Novatek Microelectronics Corp.

Min-Hung Chen*
NVIDIA

Chien-Yi Wang*
NVIDIA

Shang-Hong Lai*
National Tsing Hua University

Trista Chen
Microsoft Corp.

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. 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, to automatically learn 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 experiments on multiple public kinship recognition benchmarks. Our code is available at https://github.com/wtnthu/FaCoR.*

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 [6] 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 [22], automatic album organization [41], child adoption [35], and social media applications [7]. 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 an 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.

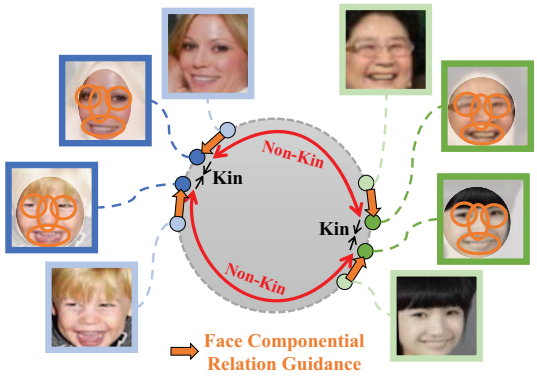

Figure 1. Our method uses face components as clues and guides the training with the relation of facial image pairs, where the relation estimation for face components (darker blue and green colors) can further pull *kin* faces together and push away *non-kin* faces, improving the efficacy of contrastive learning from the original whole-face features (lighter blue and green colors).

To deal with these challenges, several traditional approaches incorporate hand-crafted features [22] with metric learning [12] to learn discriminative features. Motivated by the success of deep learning, various methods improve kinship recognition by exploiting powerful deep feature representations. CNN-Point [38] first adopts a CNN model to extract discriminative features, outperforming previous hand-crafted ones. For the extension, several CNN-based approaches [5, 23, 37] focus on designing fusion mechanisms

*Work was done during Microsoft

to integrate the features among an image pair. Recently, the supervised contrastive approach [40] 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, most of the approaches rely on heuristic designs. For example, the feature fusion approaches [36, 38] utilize several arithmetic combinations or feature concatenation to fuse the feature pair for kinship recognition. Despite the state-of-the-art performance from [40], 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 in the image pair appear similarly, then there is a higher chance that this is a *kin* pair. 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 in Fig. 1.

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 (Rel-Guide)* based on cross-attention estimation instead of heuristic tuning [40]. 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 ***Fa**ce **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 [26]. To be specific, our work outperforms the previous best method in three tasks with standard protocol by 2.7% (79.3% $\rightarrow$ 82.0%) in the kinship verification task, 0.7% (84.4% $\rightarrow$ 85.1%) in the tri-subject verification task, and 14.2% (40.0% $\rightarrow$ 54.2%) in the search and retrieval task. We also show that our FaCoRNet achieves SOTA performance on the other two widely-used kinship recognition benchmarks, KinFaceW-I and KinFaceW-II.

Our contributions are summarized as follows:

- We propose a novel ***Fa**ce **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 training.
- The proposed ***FaCoRNet*** model outperforms previous SOTA methods by large margins on multiple standard kinship recognition benchmarks.

2. Related Work

In the past few years, several kinship recognition approaches have been proposed [3, 5, 12, 13, 15, 18, 21, 22, 23, 29, 30, 32, 37, 38, 40], where most of them focus on extracting discriminative feature for each facial image. Traditional approaches include designing hand-crafted feature extractors [1, 4, 31] and metric learning [9, 10, 15] for solving similarity metrics in kinship recognition. Recently, deep learning methods make significant advances, including two main categories: *feature fusion* and *deep metric learning*.

Feature Fusion: [38] utilizes the multiple face regions as the model inputs to learn richer facial features for kinship recognition. The multi-task deep learning-based approach [5] uses seven kinship sub-classes to jointly train with the kinship labels for kin recognition. Ustc-nelslip [36] adopts a siamese network to extract features and designs three different math operations to fuse feature pairs, followed by direct concatenation with a fully-connected layer. [29] proposes an advanced knowledge-based tensor similarity extraction framework for automatic facial kinship verification that utilizes four pre-trained networks to improve the performance.

Deep Metric Learning: [11] 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 [40] utilizes supervised contrastive loss with the ArcFace pre-trained model [8] 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 in an image pair, thus achieving state-of-the-art performance for kinship recognition. [15] presents a novel cross-pair metric learning approach that introduces a k-tuplet loss. This approach effectively captures both low-order and high-order discriminative features from multiple negative pairs.

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, and find out important facial parts for kinship recognition. Moreover, our approach incorporates the face componential correlation to adapt contrastive learning automatically, without a strong reliance on heuristic designs.

3. Proposed Methods

In this work, we propose the ***Face Componential Relation Network (FaCoRNet)***, which considers the *face components* and learns 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 computes the cross-relation among a face image pair and enhances feature representations to fully exploit the symmetry of face components in the image pair. In addition, cross-layer features are mutually interacted and fused in the channel dimension by the Channel Interaction (CI) blocks (Sec. 3.1). Moreover, 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

One core question for kinship recognition is: *How to properly extract and compute the relation between face components in 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-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 $(\mathbf{X}^a, \mathbf{X}^b)$ and adopt a learnable γ-scaled residual connection to obtain the cross-attention features $(\mathbf{O}^a, \mathbf{O}^b) \in \mathbb{R}^{C\times HW}$, given by:

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

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

To effectively fuse the information from the cross-layer features, including high-level features $(\mathbf{r}^a, \mathbf{r}^b)$ and the cross-attention features $\left(\mathbf{O^a}, \mathbf{O^b}\right)$, we adopt Channel Interaction (CI) blocks that encode inter-channel relations 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 follows:

$$\mathbf{w} = \sigma\left(\mathrm{Conv}_{1\times1}\left(\delta\left(\mathrm{Conv}_{1\times1}(\hat{\mathbf{x}})\right)\right)\right), \tag{3}$$

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 to the CI block, where the elements in $\hat{\mathbf{x}}$ are multiplied element-wise with their corresponding weights to produce a set of weighted feature values $\mathbf{w}\hat{\mathbf{x}}$. Finally, the outputs of the FaCoR module $(\mathbf{x^a_{out}}, \mathbf{x^b_{out}})$ are generated by fusing the information of cross-layer features via the Channel Interaction blocks as follows:

$$\left(\mathbf{x^a_{out}}, \mathbf{x^b_{out}}\right) = \left(\mathbf{CI}\left(\mathbf{CI}\left(\mathbf{O^a}\right) \,||\, \mathbf{r}^a\right), \mathbf{CI}\left(\mathbf{CI}\left(\mathbf{O^b}\right) \,||\, \mathbf{r}^b\right)\right), \tag{4}$$

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

3.2. Relation-Guided Contrastive Learning

Contrastive learning [2, 16] 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 [40] approach learns more robust features in kinship recognition, achieving 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 close. Otherwise, the feature representations of non-kin relations in feature space are far apart.

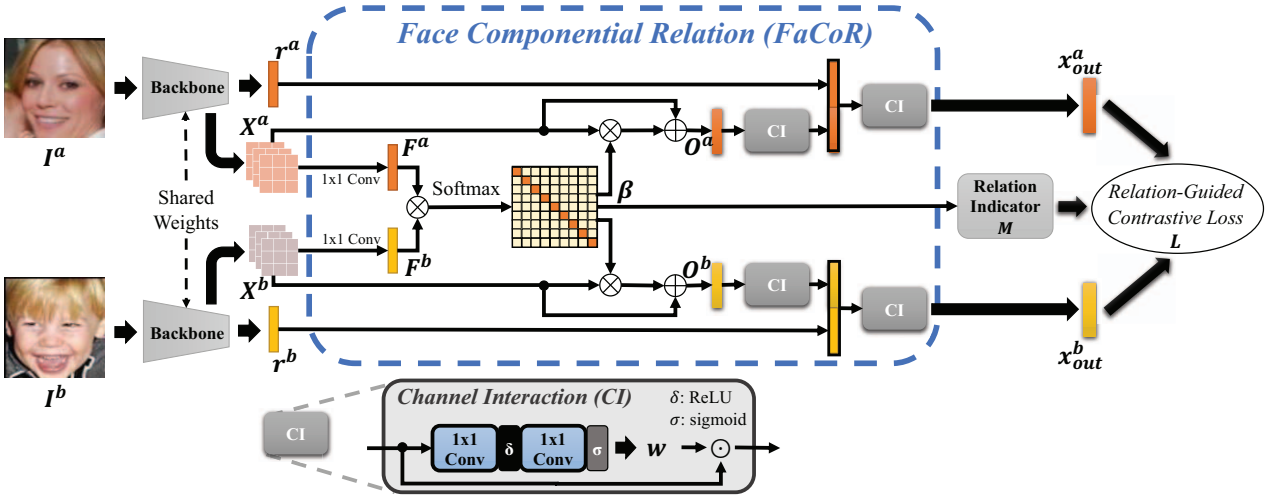

Figure 2. An overview of the proposed *Face Componential Relation Network (FaCoRNet)* consisting of a backbone and the Face Componential Relation (FaCoR) module (Sec. 3.1), trained with the Relation-Guided Contrastive Loss L (Sec. 3.2). In the FaCoR module, we compute the cross-attention features $\mathbf{O}$ embedded with face componential relation and then fuse the information from high-level features $\mathbf{r}$ via Channel Interaction (CI) blocks. During training, the attention map $\boldsymbol{\beta}$ is adopted as the guidance to learn discriminative representations.

For the standard contrastive learning, given N positive samples (x_i, y_i), the contrastive loss L is given by:

$$L = \frac{1}{2N}\sum_{i=1}^{N}(L_c(x_i, y_i) + L_c(y_i, x_i)) \tag{5}$$

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

$$L_c(x_i, y_i) = -log\frac{e^{sim(x_i,y_i)/\tau}}{\sum_{j=1}^{N}\mathbb{1}_{[j\neq i]}(e^{sim(x_i,x_j)/\tau} + e^{sim(x_i,y_j)/\tau})}, \tag{6}$$

where $\mathbb{1}_{[j\neq i]} \in \{0, 1\}$ represents an indicator function that evaluates to 1 iff $j \neq i$, the negative samples are generated by incorporating positive from different kinship categories (i.e., (x_i, x_j)), and $sim(x, y)$ is the cosine similarity operation between x and y.

However, the kinship recognition performance of contrastive learning is sensitive to hyper-parameter τ [40], which controls the degree of penalty for hard samples. To solve this problem, we propose the *Relation-Guided Contrastive Loss (Rel-Guide)* with a relation indicator $\mathbf{M}$, which guides the contrastive loss with the cross-attention estimation instead of heuristic tuning, as shown in Fig. 2.

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 τ in Eq. 7 as:

$$L_c(x_i, y_i) = -log\frac{e^{sim(x_i,y_i)/\psi}}{\sum_{j=1}^{N}\mathbb{1}_{[j\neq i]}(e^{sim(x_i,x_j)/\psi} + e^{sim(x_i,y_j)/\psi})}, \tag{7}$$

where $\psi = \mathbf{M}(\boldsymbol{\beta})/s$, s is the scale value, and we adopt the global sum pooling operation as the relation indicator $\mathbf{M}$. We refer to the contrastive learning approach [40], the hyper-parameter τ mostly lies in the range of 0.08-0.1 for stable training. Therefore, the scale value s in Rel-Guide is set to let the value learned adaptively within this range. In our FaCoRNet, the feature pair (x, y) in the loss function L_c uses the output feature pairs $(\mathbf{x^a_{out}}, \mathbf{x^b_{out}})$ from Face Componential Relation to calculate loss for updating the model.

For the inference, we follow Contrastive [40] to extract the final outputs $(\mathbf{x^a_{out}}, \mathbf{x^b_{out}})$ from Eq. 4 to calculate the cosine similarity, and then predict whether there is a kinship relation between them with thresholding.

4. Experiments

4.1. Datasets and Evaluation

The compared methods are trained and tested on three publicly available kinship recognition datasets: **Families in the Wild (FIW)** [26], and **KinFaceW-I and II** [22].

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

Method	BB	SS	SIBS	FD	MD	FS	MS	AVG.
(a) Pre-trained model: ArcFace [8]								
Stefhoer† [13]	0.660	0.650	0.760	0.770	0.770	0.800	0.780	0.740
DeepBlueAI† [23]	0.770	0.770	0.750	0.740	0.750	0.810	0.740	0.760
Ustc-nelslip† [36]	0.750	0.740	0.720	0.760	0.750	0.820	0.750	0.760
Vuvko† [30]	0.800	0.800	0.770	0.750	0.780	0.810	0.740	0.780
Contrastive [40]	0.803	0.829	0.794	0.753	0.803	0.823	0.751	0.793
FaCoRNet (Ours)	**0.820**	**0.833**	**0.810**	**0.773**	**0.804**	**0.826**	**0.788**	**0.806**
(b) Pre-trained model: AdaFace [17]								
Contrastive [40] (naive)	0.630	0.776	0.731	0.663	0.687	0.736	0.687	0.728
Contrastive [40]	0.821	0.831	0.798	0.766	0.806	0.828	0.767	0.802
FaCoRNet (Ours)	**0.832**	**0.836**	**0.824**	**0.795**	**0.818**	**0.848**	**0.802**	**0.820**

Table 1. The state-of-the-art performance comparison of *Kinship Verification* on FIW dataset by two pre-trained backbones: (a) ArcFace [8] and (b) AdaFace [17]. The best and second results in each column are in **bold** and underline, respectively. †The results are from [26].

types. All face images are cropped to the size of 112×112 with face detection and alignment in training and testing by MTCNN [39]. 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. The evaluation of FIW comprises three tasks: 1) *Kinship Verification (one-to-one)*: verify the kinship relationship to predict whether a pair of individuals are blood relatives; 2) *Tri-Subject Verification (one-to-two)*: the goal is to determine whether a child is related to a pair of parents; 3) *Search and Retrieval (many-to-many)*: the goal is to find images in the gallery (31,787 images) that are most likely to have a kinship with the probe (190 families). For evaluation, we adopt cosine similarity and thresholding to calculate accuracy according to the FIW benchmark [26].

The **KinFaceW-I and II** datasets are two widely-used kinship datasets for evaluation, which include 4 kin relationship types include: Father-Son (FS), Father-Daughter (FD), Mother-Son (MS), and Mother-Daughter (MD). The KinFaceW-I dataset contains 134 (FS), 156 (FD), 127 (MS), and 116 (MD) pairs of parent-child facial images. The KinFaceW-II dataset consists of 250 pairs of facial images for each kinship relation. For evaluation, we adopt the five-fold cross-validation in the experiments following the standard protocol in GKR [19].

4.2. Implementation Details

For experiments, we select 103784 positive and negative image pairs overall without non-aligned images for the training phase and follow the evaluation protocols as detailed in [27] 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 FaCoRNet against several existing methods by using ArcFace [8] 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 [17], 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 as 50, and the models are trained for 50 epochs. The scale value s in Relation-Guided Contrastive Loss is set to 500 for stable training.

4.3. Experimental Results

4.3.1 Comparison to SOTA Methods

We first evaluate kinship recognition performance on the FIW dataset for the three tasks given in the standard protocol, and compare our method with several state-of-the-art methods including stefhoer [13], DeepBlueAI [23], Vuvko [30], Ustc-nelslip [36], and Contrastive [40].

Table 1 compares the *Kinship Verification (one-to-one)* 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 [8] as the pre-trained model, our FaCoRNet outperforms previous leading methods Ustc-nelslip, Vuvko, and Contrastive by 4.6 percent ($0.760 \rightarrow 0.806$), 2.6 percent ($0.780 \rightarrow 0.806$), and 1.3 percent ($0.793 \rightarrow 0.806$), respectively, as shown in Table 1(a). Then a question arises: *Do advanced face recognition models benefit kinship recog-*

nition? To answer this, we adopt Contrastive [40] as the strong baseline and exploit AdaFace [17] 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, as shown in Table 1(b).

Table 2 compares the *Tri-Subject Verification (one-to-two)* performance by using the AdaFace to be the pre-trained model. The results demonstrate that the average accuracy from our FaCoRNet outperforms previous leading methods DeepBlueAI, Ustc-nelslip, and Contrastive by 8.1 percent (0.770 → 0.851), 6.1 percent (0.790 → 0.851), and 0.7 percent (0.844 → 0.851), respectively.

In the *Search and Retrieval (many-to-many)* task, the problem is transformed into a many-to-many verification task, significantly increasing the task difficulty. We apply our proposed FaCoRNet model from the kinship verification task (i.e., the pre-trained model from AdaFace and train from the verification task). Table 3 demonstrates that the search and retrieval performance from our FaCoRNet significantly outperforms current SOTA methods Vuvko, Ustc-nelslip, and Contrastive by 16.2 percent (0.390 → 0.542), 31.2 percent (0.230 → 0.542), and 14.2 percent (0.400 → 0.542), respectively. These results show that the extracted face componential relation information by our proposed FaCoRNet substantially benefits the challenging many-to-many task.

To summarize, our FaCoRNet with the proposed training scheme significantly outperforms the SOTA methods on all three tasks in the FIW benchmark, achieving new SOTA results.

We also evaluate the kinship verification performance of our method on two widely-used databases: KinFaceW-I [22] and KinFaceW-II [22], and compare our proposed FaCoRNet with several state-of-the-art methods including MNRML [22], MNRML [22], DMML [35], CNN-Basic [38], CNN-Point [38], D-CBFD [34], WGEML [20], GKR [19] and Contrastive [40]. Table 4 compares the kinship verification performance by using the pre-trained ResNet-18 in various methods. The result shows that the kinship verification accuracy from our proposed method is comparable to or higher than those achieved by the other methods.

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 5 (b). More specifically, we propose a quality-filtered protocol, where we select high-quality training and testing face images with SER-FIQ quality scores [33] larger than 0.5. The results demonstrate that the average accuracy of FaCoRNet are significantly higher than the baseline (i.e., Contrastive). This

Method	FMD	FMS	AVG.
Stefhoer† [13]	0.720	0.740	0.730
DeepBlueAI† [23]	0.760	0.770	0.770
Ustc-nelslip† [36]	0.780	0.800	0.790
Contrastive [40]	0.824	**0.860**	0.844
FaCoRNet (Ours)	**0.841**	0.857	**0.851**

Table 2. The state-of-the-art performance comparison of *Tri-subject Verification* on FIW dataset. The best and second results are in **bold** and underline, respectively. †The results are from [26].

Method	Rank@5	AVG.
Vuvko† [30]	0.600	0.390
DeepBlueAI† [23]	0.320	0.190
Ustc-nelslip† [36]	0.380	0.230
Contrastive [40]	0.600	0.400
FaCoRNet (Ours)	**0.668**	**0.542**

Table 3. The state-of-the-art performance comparison of *Search and Retrieval* on FIW dataset. The best and second results are in **bold** and underline, respectively. †The results are from [26].

Method	FS	FD	MS	MD	AVG.
KinFaceW-I					
MNRML† [22]	0.725	0.665	0.662	0.720	0.699
DMML† [35]	0.745	0.695	0.695	0.755	0.723
CNN-Basic† [38]	0.757	0.708	0.734	0.794	0.748
CNN-Point† [38]	0.761	0.718	0.780	0.841	0.775
D-CBFD† [34]	0.790	0.742	0.754	0.773	0.785
WGEML† [20]	0.785	0.739	0.806	0.819	0.787
GKR† [19]	0.795	0.732	0.780	**0.862**	0.792
Contrastive [40]	**0.799**	0.805	0.835	0.780	0.805
FaCoRNet (Ours)	**0.799**	**0.818**	**0.839**	0.806	**0.815**
KinFaceW-II					
MNRM†L [22]	0.769	0.743	0.774	0.776	0.765
DMML† [35]	0.785	0.765	0.785	0.795	0.783
CNN-Basic† [38]	0.849	0.796	0.883	0.885	0.853
CNN-Point† [38]	0.894	0.819	0.899	0.924	0.884
D-CBFD† [34]	0.810	0.762	0.774	0.793	0.785
WGEML† [20]	0.886	0.774	0.834	0.816	0.828
GKR† [19]	**0.908**	0.860	**0.912**	**0.944**	**0.906**
Contrastive [40]	0.852	0.898	**0.912**	0.890	0.888
FaCoRNet (Ours)	0.886	**0.922**	0.910	0.900	**0.906**

Table 4. Verification accuracy of different methods on KinFaceW-I and KinFaceW-II datasets. The best and second results are in **bold** and underline, respectively. †The results are from [19].

Method	BB	SS	SIBS	FD	MD	FS	MS	AVG.
(a) Standard Protocol								
Contrastive [40]	0.803	0.829	0.794	0.753	0.803	0.823	0.751	0.793
FaCoRNet (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 [40]	0.800	0.817	0.772	0.739	0.784	0.836	0.786	0.792
FaCoRNet (Ours)	**0.836**	**0.838**	**0.784**	**0.784**	**0.842**	**0.862**	**0.815**	**0.826**

Table 5. Performance comparison of kinship on FIW dataset by using AdaFace to be the pre-trained model 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.

trend is similar to the standard protocol as shown in Table 5 (a), but the improvement from our method over the baseline is even more obvious (0.792 $\rightarrow$ 0.826).

Intuitively, using high-quality face images as training and testing data would improve overall accuracy. However, the accuracy of Contrastive [40] 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 the 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 $\rightarrow$ 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. 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. Finally, the results also demonstrate that our FaCoRNet outperforms the SOTA method by a large margin in all kin categories.

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 in *Kinship Verification*, *Tri-subject Verification*, and *Search and Retrieval* tasks on the FIW dataset, corresponding to task 1, task 2, and task 3, respectively. The two main components in our proposed FaCoRNet are the **FaCoR** module and the **Rel-Guide** loss function. *FaCoR* learns to discover important facial components to benefit kinship-related tasks, and *Rel-Guide* facilitates training for further improvement, especially for more challenging tasks. As shown in Table 6, FaCoR consistently improves the baseline for all FIW [28, 26] tasks. For the most arduous *Search and Retrieval* task, Rel-Guide can further boost the accuracy. This was achieved by using *Relation-Guided Contrastive Loss (Rel-Guide)* with a relation indicator, which automatically estimate τ value in a certain range in the contrastive loss instead of heuristic tuning of fixed τ, particularly on the challenging task 3.

Contrastive	FaCoR	Rel-Guide	task 1	task 2	task 3
✓			0.793	0.844	0.400
✓	✓		0.803	0.848	0.511
✓	✓	✓	**0.806**	**0.851**	**0.542**

Table 6. Component analysis of FaCoRNeton the FIW dataset, including *Kinship Verification*, *Tri-subject Verification*, and *Search and Retrieval* tasks, corresponding to task 1, task 2, and task 3, respectively.

Face Quality Analysis: We perform an experiment to evaluate accuracy of applying different methods to face images with different image qualities. We use SER-FIQ [33] 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 7. The results show that in low-quality cases (i.e., the quality scores$<$0.4), the overall recognition accuracy is lower than those 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 different levels of face qualities.

Face-Pair Quality Score	Contrastive [40]	**FaCoRNet (Ours)**
0-0.2	0.749	**0.794**
0.2-0.4	0.782	**0.820**
0.4-0.6	0.813	**0.843**
0.6-0.8	0.803	**0.821**
0.8-1	0.793	**0.824**
AVG.	0.793	**0.820**

Table 7. 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 Verification: We also evaluate face verification performance with two databases: LFW [14] and AgeDB-30 [24], and compare our proposed FaCoRNet with two methods including the state-of-the-art Contrastive [40] approach

Method	ACC.	AUC
(a) LFW dataset		
ArcFace-MS1M† [8]	0.998	0.999
Contrastive [40]	0.993	0.998
FaCoRNet (Ours)	**0.995**	0.999
(b) AgeDB-30 dataset		
ArcFace-MS1M† [8]	0.980	0.991
Contrastive [40]	0.965	0.989
FaCoRNet (Ours)	**0.970**	0.989

Table 8. The results of *Face Verification* comparison on LFW and AgeDB-30 dataset. †The upper-bound of the face verification.

and the regular face recognition model (Arcface-MS1M [8]). The regular face recognition model of Arcface-MS1M is trained by using the MS1MV3 [25] database, which represents the upper bound of face recognition. As shown in Table 8, the results show that the face verification performance of our proposed method is significantly higher than the Contrastive scheme. It demonstrates that our proposed method not only learns the discriminative feature of kinship recognition but also retains the identity information.

Visualization: We also perform visual analysis on the latent features learned by our proposed model, as shown in Fig. 3. We select 5 families from the FIW validation set and visualize the distribution by using t-SNE. We observe that members of the same family are close to each other, while there are gaps between members of different families. It indicates the discriminative ability of our kinship model.

Besides, Fig. 4 visualizes the cross-attention map β produced by our proposed model from Eq. 1. The visualization results demonstrate that our proposed method FaCoRNet attains higher values in the face components, particularly in the regions of the eye, nose, and mouth where there is a high degree of similarity, and where the covering is skipped. These findings substantiate that our FaCoRNet method can effectively learn the relevance of the face components in image pairs through cross-attention estimation, and can adaptively acquire critical face components for accurate kinship recognition. Different from many existing works which also utilize relation information between face components, we do not manually define partitions of the face regions. Instead, our FaCoR and supervised loss let the network automatically focus on the important regions that can benefit kinship tasks by modeling the relation between regions. Therefore, the focused regions are not always identical but could be different areas (e.g., cheek, forehead), depending on the input face pair.

5. Conclusion and Future Work

In this paper, we propose a novel ***Fa**ce **Co**mponential **R**elation **Net**work (**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. Experimental results show that our method achieves SOTA performance on multiple kinship recognition benchmarks, including the FIW benchmark. Moreover, for practical kinship recognition protocol, FaCoRNet also outperforms the SOTA methods by large margins. We believe that FaCoRNet 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.

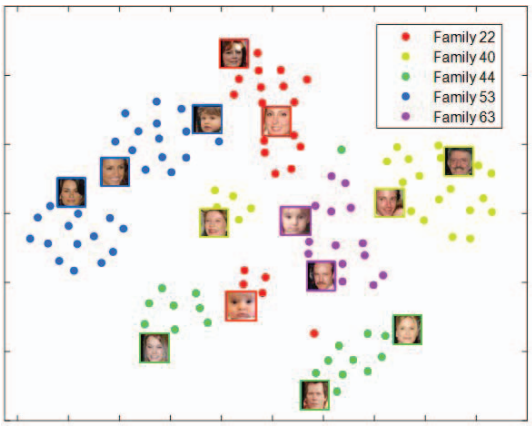

Figure 3. Visual analysis of the learned feature with t-SNE.

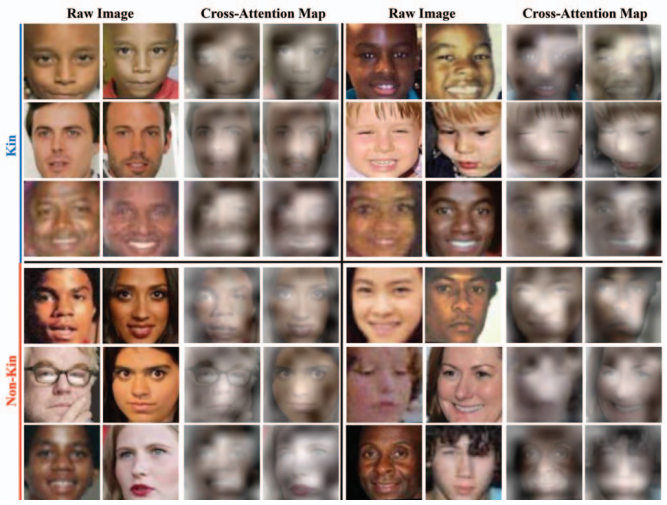

Figure 4. Illustration of Cross-Attention map includes kin (first to third rows) and non-kin (fourth to sixth rows) cases: the left side shows the raw image pairs and the right side shows the visualization of the cross-attention maps.

References

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

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[3] Xiaopan Chen, Changlong Li, Xiaoke Zhu, Liang Zheng, Ya Chen, Shanshan Zheng, and Caihong Yuan. Deep discriminant generation-shared feature learning for image-based kinship verification. *Signal Processing: Image Communication*, 101:116543, 2022. 2

[4] 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*, pages 3554–3561, 2013. 2

[5] Eran Dahan and Yosi Keller. A unified approach to kinship verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2851–2857, 2020. 1, 2

[6] 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. 1

[7] 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*, pages 1757–1764, 2014. 1

[8] 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*, pages 4690–4699, 2019. 2, 5, 8

[9] Hamdi Dibeklioglu. Visual transformation aided contrastive learning for video-based kinship verification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2459–2468, 2017. 2

[10] 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. 2

[11] 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*, pages 1590–1598, 2017. 2

[12] 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. 1, 2

[13] 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)*, pages 863–867. IEEE, 2020. 2, 5, 6

[14] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 7

[15] Sheng Huang, Jingkai Lin, Luwen Huangfu, Yun Xing, Junlin Hu, and Daniel Dajun Zeng. Adaptively weighted k-tuple metric network for kinship verification. *IEEE Transactions on Cybernetics*, 2022. 2

[16] 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. 3

[17] 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*, pages 18750–18759, 2022. 5, 6

[18] Wanhua Li, Shiwei Wang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Meta-mining discriminative samples for kinship verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16135–16144, 2021. 2

[19] Wanhua Li, Yingqiang Zhang, Kangchen Lv, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Graph-based kinship reasoning network. In *2020 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2020. 5, 6

[20] Jianqing Liang, Qinghua Hu, Chuangyin Dang, and Wangmeng Zuo. Weighted graph embedding-based metric learning for kinship verification. *IEEE Transactions on Image Processing*, 28(3):1149–1162, 2018. 6

[21] 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)*, pages 1–8. IEEE, 2021. 2

[22] 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. 1, 2, 4, 6

[23] 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)*, pages 868–871. IEEE, 2020. 1, 2, 5, 6

[24] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 7

[25] Federico Pernici, Matteo Bruni, Claudio Baecchi, and Alberto Del Bimbo. Maximally compact and separated features with regular polytope networks. In *CVPR Workshops*, pages 46–53, 2019. 8

[26] 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. 2, 4, 5, 6, 7

[27] 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. 5

[28] Joseph P Robinson, Ming Shao, Yue Wu, Hongfu Liu, Timothy Gillis, and Yun Fu. Visual kinship recognition of families in the wild. *TPAMI*, 2018. 7

[29] I Serraoui, Oualid Laiadi, Abdelmalik Ouamane, Fadi Dornaika, and Abdelmalik Taleb-Ahmed. Knowledge-based tensor subspace analysis system for kinship verification. *Neural Networks*, 151:222–237, 2022. 2

[30] Andrei Shadrikov. Achieving better kinship recognition through better baseline. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 872–876. IEEE, 2020. 2, 5, 6

[31] Gowri Somanath and Chandra Kambhamettu. Can faces verify blood-relations? In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 105–112. IEEE, 2012. 2

[32] Chaohui Song and Haibin Yan. Kinmix: A data augmentation approach for kinship verification. In *2020 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2020. 2

[33] 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*, pages 5651–5660, 2020. 6, 7

[34] Haibin Yan. Learning discriminative compact binary face descriptor for kinship verification. *Pattern Recognition Letters*, 117:146–152, 2019. 6

[35] 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. 1, 6

[36] 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)*, pages 892–899. IEEE, 2020. 2, 5, 6

[37] 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)*, pages 1–8. IEEE, 2021. 1, 2

[38] 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. 1, 2, 6

[39] 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. 5

[40] 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)*, 2021. 2, 3, 4, 5, 6, 7, 8

[41] 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*, pages 725–728, 2012. 1