Acta Cybernetica **26** (2023) 215–241.

Dual Convolutional Neural Network Classifier with Pyramid Attention Network for Image-Based Kinship Verification*

Reza Fuad Rachmadi[ab], I Ketut Eddy Purnama[ac],
Supeno Mardi Susiki Nugroho[de], and Yoyon Kusnendar Suprapto[df]

Abstract

A family is the smallest entity that formed the world with specific characteristics. The characteristics of a family are that the member can/may share some similar DNA and leads to similar physical appearances, including similar facial features. This paper proposed a dual convolutional neural network (CNN) with a pyramid attention network for image-based kinship verification problems. The dual CNN classifier is formed by paralleling the FaceNet CNN architecture followed by family-aware features extraction network and three final fully-connected layers. A channel-wise pyramid attention network is added after the last convolutional layers of FaceNet CNN architecture. The family-aware features extraction network is used to learn family-aware features using the SphereFace loss function. The final features used to classify the kin/non-kin pair are joint aggregation features between the pyramid attention features and family-aware features. At the end of the fully connected layer, a softmax loss layer is attached to learn kinship verification via binary classification problems. To analyze the performance of our proposed classifier, we performed experiments heavily on the Family in The Wild (FIW) kinship verification dataset. The FIW kinship verification dataset is the largest dataset for kinship verification currently available. Experiments of the FIW dataset show that our proposed classifier can achieve the highest average accuracy of 68.05% on a single classifier scenario and 68.73% on an ensemble classifier scenario which is comparable with other state-of-the-art methods.

*This research is partially supported by Indonesia Minister of Education, Culture, and Higher Education Grant (under PDUPT scheme) number 930/PKS/ITS/2019 and 1224/PKS/ITS/2020.

[a]Dept. of Computer Engineering and University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS), Institut Teknologi Sepuluh Nopember, Indonesia

[b]E-mail: fuad@its.ac.id, ORCID: 0000-0001-9101-5598

[c]E-mail: ketut@te.its.ac.id, ORCID: 0000-0002-7438-7880

[d]Dept. of Computer Engineering, Institut Teknologi Sepuluh Nopember, Indonesia

[e]E-mail: mardi@its.ac.id, ORCID: 0000-0001-8109-6136

[f]E-mail: yoyon@te.its.ac.id, ORCID: 0000-0003-3149-5088

DOI: 10.14232/actacyb.296355

Keywords: dual CNN classifier, image-based kinship verification, Deep Metric Learning, Channel-wise Pyramid Attention Network

1 Introduction

Humans are unique species in the universe that discriminate by visual appearances, including human faces, fingerprints, retina patterns, and gait. All of those visual appearances are widely used as biometric authentication features of identity. The human faces are a little bit special due to the visual appearances that can be descent from parents or grandparents and can be used to analyze the kinship relationship among people. In this modern era, the camera sensor is widely used to capture images. Many of those images were uploaded to the internet, including photos with human faces and family photos. In recent years, several kinship relationship datasets were formed by researchers to support the development of image-based kinship relationship problems, including KinFaceW-I [26, 27], KinFaceW-II [26, 27], KFVW (Kinship Face Video in The Wild) [54], Cornell KinFace [12], Tri-Subject Kinship [31] and FIW (Family in The Wild) [37, 50, 35]. The dataset is usually formed by crawling well-known families' photos on the web, including actresses and the royal family with clear kinship relationships between the family members.

One of the technologies that provide an opportunity to develop image-based kinship verification problems is the evolution of deep learning methods widely used after Krizhevsky et al. [17] won the ILSVRC 2012 challenges by using a convolutional neural network classifier. After 2012, deep learning is constantly used for a lot of problems and applications, from computer science to remote sensing applications. There are several deep learning approaches for image-based kinship verification problems, including the one described in [20, 21, 11, 50, 8, 32, 35, 33, 30, 36, 53].

In this paper, we proposed a dual convolutional neural network (CNN) classifier with joint features aggregation and a pyramid attention network for image-based kinship verification problems. Our proposed classifier was formed by paralleling the FaceNet CNN architecture [40] and adding two subnetworks, one for family-aware features extraction and one for kin/non-kin classification. Our contributions can be listed as follows.

- We investigated a dual CNN classifier with joint features aggregation for image-based kinship verification problems. The experiments are heavily performed using the FIW dataset [37, 50, 35], which is considered the largest kinship verification dataset currently available.
- We investigated the combination of our proposed classifier with a channel-wise pyramid attention network. The attention network described by Zhao and Wu [58] is adopted with our proposed classifier. Experiments on the FIW dataset show that adding a channel-wise pyramid attention network can improve the classifier's performance.

- For further analysis, we also investigated our proposed classifier with a subset of the FIW dataset, including RFIW'17 [38] and RFIW'18 [35]. The subset of the FIW dataset is used for the competition, which can be compared side-by-side with other methods in the competition.

The rest of the paper is organized as follows. Section 2 discussed several related works on image-based kinship relationship analysis. Our proposed classifier is described in section 3, follows by results and discussion in section 4. Finally, we conclude the experiments in the last section.

2 Related Work

In recent years, there are several works on image-based kinship relationship analysis, including those described in [8, 49, 33, 39, 56, 19]. Dawson et al. [8] reported a performance comparison FSP (From-Same-Photo) classifier on several kinship verification datasets. The FSP classifier is trained on the same photo dataset instead of kinship verification data. The results show that the performance is very good on some kinship verification datasets (some achieved around 90%). Dawson et al. [8] conclude that some kinship verification datasets are not suitable for model development because a lot of the data is taken from the same photo.

Robinson et al. [39] described the RFIW 2020 challenges results with three tasks: kinship verification, tri-subject verification, and search & retrieval of missing children. To create a baseline performance, Robinson et al. [39] use SphereFace CNN classifier [24] which proved to produce high accuracy on face recognition tasks. The baseline performances of the SphereFace classifier are 64% on kinship verification tasks, 68% on tri-subject kinship verification, and mAP of 0.02 on missing children search & retrieval tasks.

Yu et al. [56] proposed a deep fusion siamese network for kinship verification problems. The deep siamese network is used to extract the features of two faces input. The features are fed into a features fusion network before classifying using fully connected layers with a sigmoid activation function at the network's end. Yu et al. [56] perform experiments using several different features fusion mechanisms and two different loss functions, including BCE (Binary Cross Entropy) loss and focal loss. Experiments on RFIW 2020 dataset show that the proposed classifier achieves an average accuracy of 76% on kinship verification problems and 79% on tri-subject kinship verification problems.

A combination of the Young Generation Model with Sparse Discriminative Metric Loss (SDM-Loss) was proposed by Wang et al. [49] for kinship verification problems, especially for parents-child and grandparents-grandchild kinship. The model is based on StarGAN CNN architecture described by Choi et al. [5] and modified the loss with SDM-Loss. Experiments on 5-folds FIW dataset show that ResNet+SDMLoss with an additional young generation model can achieve an average accuracy of 68.68% with siblings and 69.47% without siblings kinship. The testing is divided into two protocols because the proposed classifier uses a young generation model that may not work properly when combined with siblings kinship that has lower different ages than parent-child or grandparents-grandchild kinship.

Laiadi et al. [19] use Multilinear Side-Information based Discriminant Analysis integrating Within Class Covariance Normalization (MSIDA+WCCN) to train a model for image-based kinship verification problems. The features used by the model are extracted from the fc6 and fc7 layers of four VGG-based CNN that are trained using the ImageNet dataset. The final decision is decided using a simple cosine similarity score between features extracted from two faces using the MSIDA+WCCN model. The proposed model was tested using the KinFaceW dataset and achieved an average accuracy of 87.65% and 87% on the KinFaceW-I and KinFaceW-II datasets.

3 Proposed Classifier

This section describes our proposed classifier, which consists of two different things, the dual CNN classifier with family-aware features and channel-wise pyramid attention network. Figure 1 shows the diagram of our proposed classifier.

3.1 Dual Convolutional Neural Network

The dual CNN classifier of our proposed classifier is formed by paralleling FaceNet CNN architecture which will process for each face image pair. An additional family-aware features extraction network is attached at the end of the classifier, which is adapted from [33]. We use joint features aggregation between pyramid features and family-aware features to improve the classifier's performance. Those joint features aggregation networks proved can improve the classifier's performance in some tasks, including super-resolution tasks [22] and remote sensing image classification [28]. Unlike the dual CNN classifier used in [33], the backbone of our dual CNN classifier weights is not frozen but updated in the training process with a 0.001 times lower learning rate comparing with a fully connected and pyramid attention network. We use three different loss functions that can be computed as follows.

$$L = L_k + \alpha(L_{f1} + L_{f2}) \tag{1}$$

with L_k is the loss function of kin/non-kin classification loss, L_{f1} and L_{f2} is the loss function for learning family-aware features, and α is the contributing factor to the final loss value. We use $\alpha > 1$ for the training process, which will let the classifier learn the family-aware features strongly.

To learn the family-aware features, we use two different deep metric learning widely used for face recognition tasks, including SphereFace [24] and Center Loss [51]. Deep metric learning can be divided into two categories, euclidean metric-based loss [43, 42, 40, 51, 13] and cosine metric-based loss [25, 47, 24, 48, 9]. The SphereFace is deep metric learning that cosine metric-based loss function, which

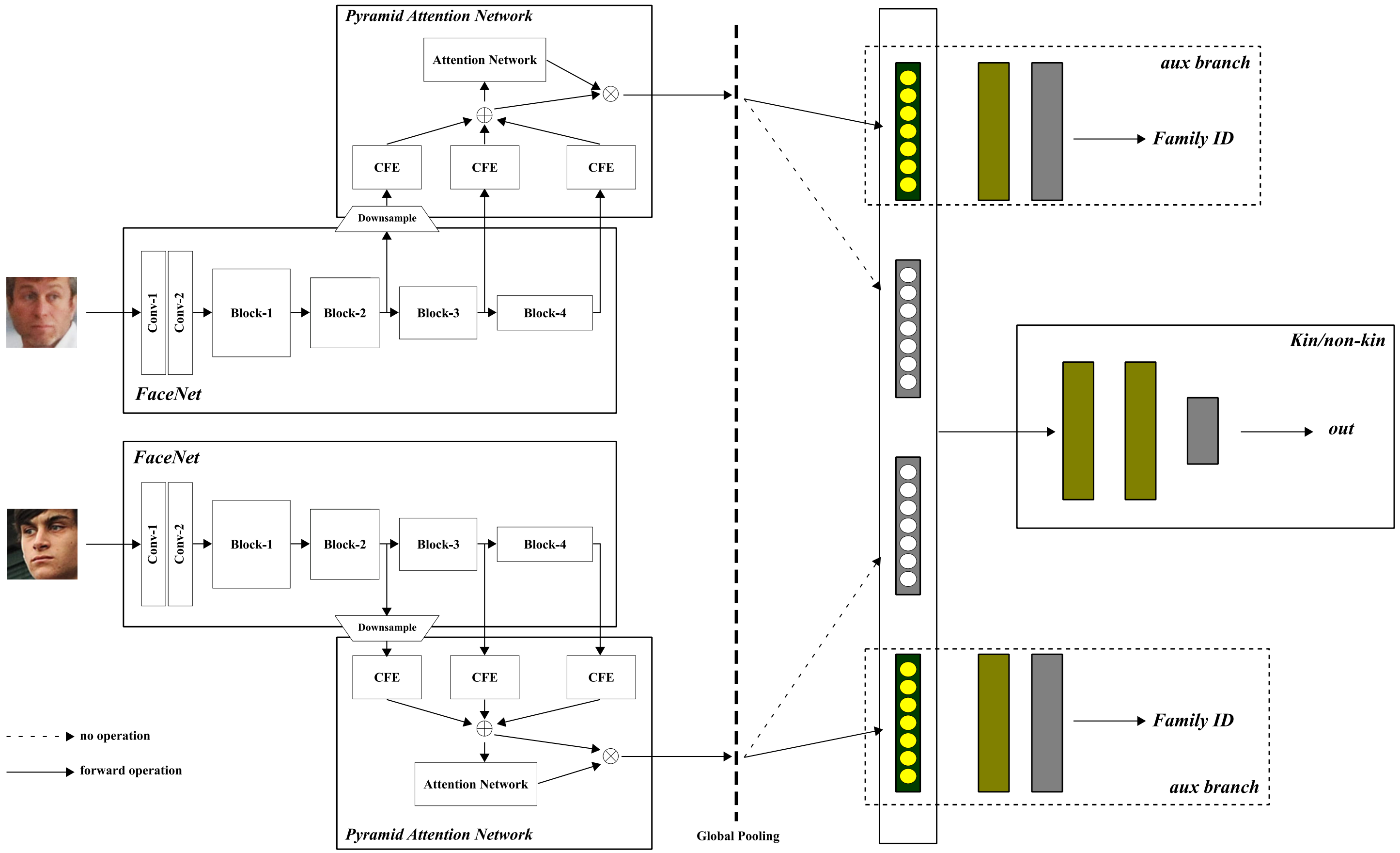


Figure 1: Diagram of our proposed classifier with joint features aggregation and channel-wise pyramid attention network. The face images are taken from the FIW dataset [37, 50, 35].

can be computed as follows.

$$L_a = \frac{1}{N}\sum_{i=1}^{N} - \log\left(\frac{e^{||\mathbf{x}_i||\psi(\theta_{c_i,i})}}{e^{||\mathbf{x}_i||\psi(\theta_{c_i,i})} + f_s(c_i)}\right) \tag{2}$$

$$f_s(c_i) = \sum_{j \neq c_i} e^{||\mathbf{x}_i|| \cos(\theta_{c_i,i})} \tag{3}$$

with $\psi(\theta_{c_i,i})$ defined as $\psi(\theta_{c_i,i}) = (-1)^k \cos(m\theta_{c_i,i}) - 2k$, $\theta_{c_i,i} \in \left[\frac{k\phi}{m}, \frac{(k+1)\phi}{m}\right]$, and $k \in [0, m-1]$. We use $m = 4$ to performs the training process as described in the original SphereFace paper [24]. The second deep metric learning used to train our proposed classifier is center loss [51]. The center loss works by minimizing the variation of the intra-class features while trying to separate the features between classes. The loss function for center loss is divided into two functions; the first loss function is used to update the center or centroid of the features, while the second loss function is used to classify the features based on their label. Let $\mathbf{x}_i$ is the extracted features of the last layer of the classifier and $\mathbf{c}_{y_i}$ is the centroid of the features of class y_i-th of the data, the loss function used for updating the center can be computed as follows.

$$L_c = \frac{1}{2}\sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{c}_{y_i}||_2^2 \tag{4}$$

The efficient way to update the centroid of the features is by analyzing all of the examples and deciding the centroid's shift based on the error produced by the examples. The process is not possible when training the classifier using the mini-batch SGD algorithm. Instead, Wen et al. [51] proposed a joint loss function between softmax and center loss that can be computed as follows.

$$L_f = L_s + \mu L_c \tag{5}$$

$$= -\sum_{i=1}^{m} \log \frac{e^{\mathbf{W}_{\mathbf{y}_i}^T \mathbf{x}_i + b_{\mathbf{y}_i}}}{\sum_j^n e^{\mathbf{W}_j^T \mathbf{x}_i + b_j}} + \frac{\mu}{2}\sum_{i=1}^{m} ||\mathbf{x}_i - \mathbf{c}_i||_2^2 \tag{6}$$

with μ is the contribution of the center loss in the final loss function, and L_s is the softmax loss function. We use $\mu = 0.008$ to performs the training process, which the original authors also recommend.

3.2 Channel-wise Pyramid Attention Network

Attention network is one type of additional network that explores the importance of features on the tasks. The attention network is widely and originally used for natural language processing problems, including that described in [2, 29, 41, 55, 45, 3, 44, 1, 10]. As time goes by, some researchers also tried to implement an attention network for the problem with an image as an input of the classifier, including that described in [46, 15, 52, 57, 58]. Zhao et al. [58] proposed a pyramid attention network for saliency detection problems. The pyramid attention network consists of two types of attention network, channel-wise attention network and spatial attention network. The channel-wise pyramid attention network computes the importantness of the features per channel, while the spatial attention network computes importantness per feature based on spatial coordinates.

Our proposed classifier adopted the channel-wise pyramid attention network (PAN) described by Zhao et al. [58] and joined the features with family-aware features [33]. Assume that $\mathbf{z} \in \mathbb{R}^{W \times H \times C}$ is the concatenation of multi-level convolutional layer outputs with $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_C]$, C is the total channel number of the features, and $\mathbf{z}_i \in \mathbb{R}^{W \times H}$ is the i-th channel of $\mathbf{z}$, the output of channel-wise attention network can be calculated as follows.

$$A_c(\mathbf{v}, W) = \sigma(f_{c_2}(\delta(f_{c_1}(\mathbf{v}, W_1)), W_2)) \tag{7}$$

with $\mathbf{v} \in \mathbb{R}^C$ is unfold version of $\mathbf{z}$, W is the parameters in the channel-wise attention network, σ is the sigmoid function, δ is the ReLU function, f_{c_1} and f_{c_2} is the fully-connected function. In our implementation, we use the PReLU function [14] instead of the ReLU which used in the original implementation. We change the activation function to match the activation function used in the backbone network (FaceNet architecture). The final features are computed by weighting the features with the output of the channel-wise attention network as follows.

$$\tilde{\mathbf{z}} = \mathbf{z} \cdot A_c(\mathbf{v}, W) \tag{8}$$

The operation is performed channel-wise multiplication using the attention weights.

The features used to calculate the channel-wise attention outputs are extracted using CFE (Context14 aware Features Extraction) module, which is also used in the original pyramid attention network paper [58]. The CFE module consists of four convolutional layers with different kernel size and dilation rates, 1×1 kernel, 3×3 kernel with dilation rates of 3, 3×3 kernel with dilation rates of 5, and 3×3 kernel with dilation rates of 7. The output of the CFE module is the combination of those four convolutional with additional batch normalization and PReLU activation functions. As shown in Figure 1, the output of convolutional blocks 2, 3, and 4 is used to extract pyramid features using the CFE module and combined it to form the final pyramid features. Features extracted from convolutional block two are downsampled to match the resolution of other features.

3.3 Face Segmentation

To ensure that the classifier only learned the appropriate face features, we applied a face parsing/face segmentation of the input faces in the preprocessing step before the training process. We use a face labeling model described in [4], which utilizes the face labeling problem described in [23] and is used to supplying the semantic segmentation for thermal-to-visible image translation using a generative adversarial network. The face parsing model produces eleven labels of face images, including

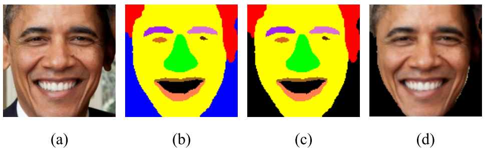


Figure 2: The results of the face parsing model that was used in our pre-processing step. (a) face image, (b) the parsing result with background, (c) the parsing result without background, and (d) the final result used in the training process.

background, left eye, right eye, left eyebrow, right eyebrow, nose, upper lips, lower lips, inside the mouth, facial skin, and hair.

In our experiments, we only take pixels that label as non-background (eyes, eyebrows, nose, lips, mouth, facial skin, and hair) for the training process. Figure 2 shows the face parsing process and removing the background labeled pixels before saved the final images for the training process (as showed in Figure 2-(d)). By using the preprocessing face images, our proposed classifier can achieve a good validation accuracy comparing without the face parsing preprocessing process. The face segmentation preprocessing process is only applied in the training process.

4 Results and Discussion

To evaluate our proposed classifier, we performed a detailed analysis using the FIW dataset [37, 50, 35] and Caffe deep learning framework [16]. We also performed experiments using only a family-aware CNN classifier [33] and several ensemble configurations. Figure 3 shows the flow of the kinship verifiation experiments for our proposed classifier.

4.1 FIW Dataset

The FIW dataset [37, 50, 35] is currently the largest kinship verification dataset and proved to be a challenging problem. The FIW dataset consists of 11,932 face images covering around 1,000 families with eleven different kinship relationship types. The eleven kinship relationship can be divided into three categories, same generation kinship (siblings, brother, and sister), first-generation kinship (mother-son, mother-daughter, father-son, and father-daughter), and second-generation kinship (grand mother-grand son, grand mother-grand daughter, grand father-grand son, and grand father-grand daughter). Figure 4 shows several examples of face images for each kinship category on the FIW dataset. As shown in Figure 4, higher generation kinship may reduce the facial features similarity which reasonable due to combination of DNA from grand parent to parent to grand child. There are several different split configurations (training and testing list) of the FIW dataset. This paper uses three different configurations, including the 5-folds configurations

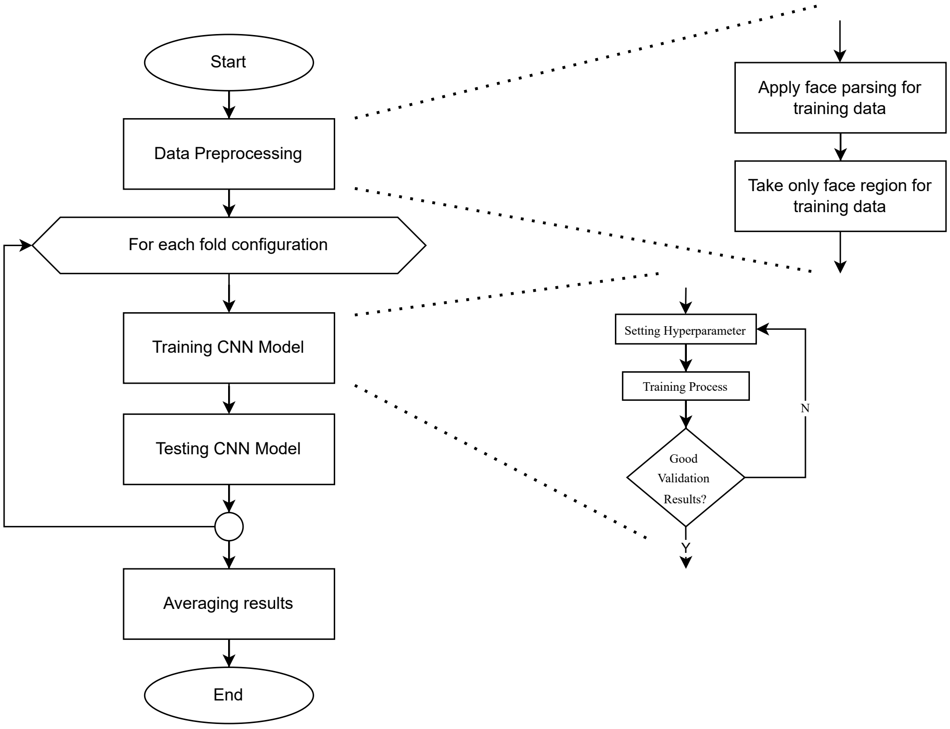


Figure 3: The main flow of our experiments using the FIW dataset.

(no overlapped family between folds), RFIW 2017 challenge, and RFIW 2018 challenge. We heavily perform the experiments using the 5-folds configuration before use RFIW 2017 and RFIW 2018 split configuration.

4.2 Experiments Setup

Implementation Detail. We use four different classifier configurations of our dual CNN classifier with a pyramid attention network. All approaches are based on FaceNet CNN architecture, and final features are constructed by combining family-aware features with pyramid attention features. Each classifier can be described as follows.

- **DFaceNet-FC512-CAtt**. Dual FaceNet classifier combined with 512 family-aware features learned using SphereFace Loss function [24] and channel-wise attention network (CAtt). The total features used for the final fully-connected layers are 896 features with 512 family-aware features and 384 features from the pyramid attention network.

- **DFaceNet-FC1K-CAtt** and **DFaceNet-FC2K-CAtt**. The classifier uses the same configuration as the DFaceNet-FC512-CAtt classifier but with a

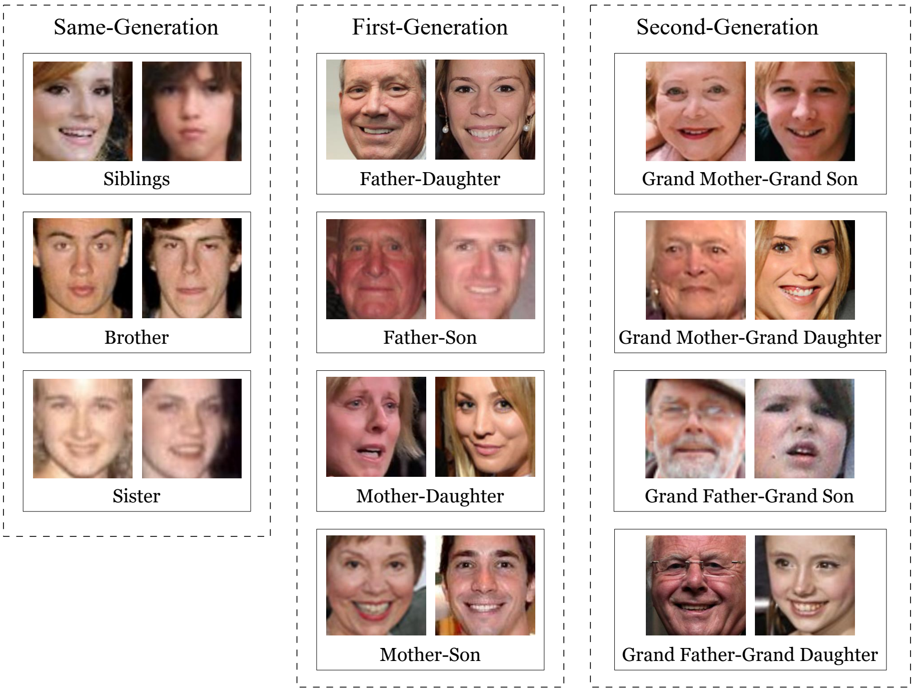


Figure 4: Kinship relationship categories in the FIW dataset and their examples pair of each category [37, 50, 35].

different number of family-aware features, 1,024 and 2,048. The total number of final features for DFaceNet-FC1K-CAtt and DFaceNet-FC2K-CAtt is 1,408 and 2,432 features, respectively.

- **DFaceNet-FC512-ASCL-CAtt**. Similar to DFaceNet-FC512-CAtt but with two different family-aware features, features learned using SphereFace Loss function and features learned using Center Loss function. The final features are 1,024 family-aware features (from two different family-aware branches) and 384 pyramid attention features.

All of the classifiers use a CPFE network with four different atrous convolutions, 1×1 kernel with dilation rate of 1 and 3×3 kernel with dilation rate of 3, 5, and 7. The CPFE network is attached after each output of blocks 2 to 4, and the pyramid features are constructed by combining the output of all CPFE networks.

Training Process. The training process is done for ten epochs using NAG (Nesterov Accelerated Gradient) training algorithm. The learning rate is initialized at 0.01 with a polynomial reducing policy and additional clipping gradient method to reduce the exploding gradient problem, especially in the first couple of epoch. We reduce the learning rate by 0.001 factor for the backbone network to preserve the classifier's ability to extract face features. The input images are resized to 120×120, followed by random cropping using 112×112 resolution and data normalization before the training process.

Testing Process. In the testing process, we use multi-resolution approaches by classifying the input image using several different input resolutions, including 115×115, 118×118, 122×122, and 128×128. Each resolution is cropped into ten different crops (center, left top, left bottom, right top, right bottom, and their respective mirror version of the crops) with a resolution of 112×112. After the pre-processing process, the testing process performs a classification using 40 crops, and the final classification score is computed by averaging the score of all crops. We performed ensemble testing by using a simple average ensemble mechanism, which proved to improve the classifier's performance by around 1-2%.

4.3 Results and Discussion

The results are divided into four different independent experiments, which are detailed discussed in each sub-section. We added one additional preliminary experiment using the FA-CNN classifier [33], which was used as the basis for our proposed classifier.

4.3.1 Preliminary Results

In the preliminary experiments, we use Dual FaceNet-FA (Family-Aware CNN) [33] with the SphereFace Loss function to learn the family-aware features. We use different training scenarios, which are not time-consuming, as reported in the original paper. The training process is done with the same hyperparameter setting as described in the experiment's setup. Lambda $\lambda = 10$ is used for the SphereFace Loss function, which in the original paper suggested choosing a small lambda value (e.g. 10 or 5) to compensate for the original softmax loss function. Table 1 shows the result of the preliminary experiments using the 5-folds FIW dataset and three different classifier configurations. As shown in Table 1, the average accuracy of the classifier is similar to the one reported in [33], although we use a different training scenario. The second-generation kinship relationship still produces the lowest accuracy due to the limited data available in the dataset.

4.3.2 5-Folds Configuration

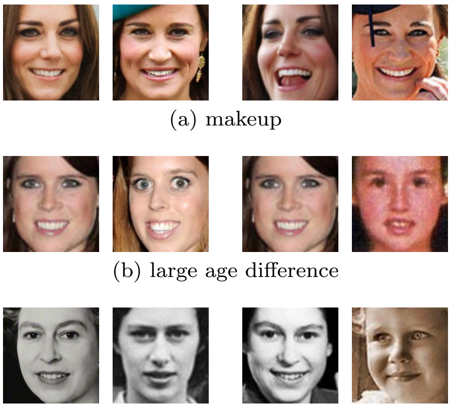
After preliminary experiments, we conducted the experiments using a Dual FaceNet classifier with family-aware features and channel-based pyramid attention network features. Four different classifiers along with ensemble configuration were used to perform the experiments. Table 2 shows the results of the Dual FaceNet classifier with family-aware features and channel-wise pyramid attention network features with average accuracy ranged from 67.80% to 68.05%. As shown in Table 2, the best performance of the single classifier is achieved using the DFaceNet-FC1K-CAtt classifier with an average accuracy of 68.05%. The second-generation kinship verification seems still the hardest case for the classifier with average accuracy

Table 1: Verification results (%) on FIW dataset for Dual FaceNet classifier using 5-fold configuration (no family overlapped between folds).

#	Method	siblings			parent-child				grandparent-grandchild				Avg
		ss	**bb**	**sibs**	**fd**	**fs**	**md**	**ms**	**gfgd**	**gfgs**	**gmgd**	**gmgs**	
1.	DFaceNet-FC512-λ10	74.8	69.0	70.3	68.8	68.1	71.8	70.2	**62.0**	**62.9**	**62.8**	64.3	**67.78**
2.	DFaceNet-FC1K-λ10	74.8	68.7	70.4	69.0	**71.9**	70.3	68.0	**62.0**	62.7	61.3	**64.5**	67.65
3.	DFaceNet-FC2K-λ10	**75.3**	**69.5**	**70.4**	**69.1**	68.4	**72.4**	**70.6**	61.9	61.8	61.2	64.1	67.75

Table 2: Verification results (%) on FIW dataset for Dual FaceNet with family-aware features and channel-based attention network using 5-fold configuration (no family overlapped between folds).

#	Method	siblings			parent-child				grandparent-grandchild				Avg
		ss	**bb**	**sibs**	**fd**	**fs**	**md**	**ms**	**gfgd**	**gfgs**	**gmgd**	**gmgs**	
1.	DFaceNet-FC512-CAtt	75.0	69.4	70.4	68.9	68.0	71.6	70.4	63.3	61.5	62.9	63.8	67.80
2.	DFaceNet-FC1K-CAtt	75.5	69.8	70.5	69.4	68.1	72.0	70.7	62.6	62.1	63.1	64.2	68.05
3.	DFaceNet-FC2K-CAtt	75.7	69.6	70.6	69.2	68.2	72.1	70.4	62.6	61.5	62.1	63.7	67.85
4.	DFaceNet-FC512-CL-CAtt	75.7	69.9	71.2	69.0	68.4	71.9	70.3	62.2	62.9	62.1	63.9	67.98
5.	Ensemble 1 + 4	75.9	70.0	71.2	69.5	68.7	72.4	71.0	**63.4**	**63.0**	62.9	63.7	68.38
6.	Ensemble 2 + 4	76.1	70.3	71.3	69.8	68.8	72.7	71.3	63.0	62.6	**63.5**	64.0	68.55
7.	Ensemble 1 + 2 + 4	76.0	**70.3**	71.4	69.9	69.0	72.8	71.4	63.2	62.5	63.4	64.0	68.59
8.	Ensemble All	**76.2**	**70.3**	**71.5**	**70.1**	**69.0**	**73.0**	**71.5**	63.4	62.7	63.5	**64.4**	**68.73**



(a) makeup

(b) large age difference

(c) color

Figure 5: Examples of the correct classification (left) and incorrect classification (right) on sister kinship relationship using same person pair.

ranged from 62% to 64%. The highest accuracy appears on sister and mother-daughter kinship type, which may be supported by the fact that sister and mother-daughter may like the same favourite makeup style and may contribute to the training process.

We take a quick analysis for the sister kinship relationship category, and Figure 5 shows the pair that correctly classify (left side) and incorrectly classify (right side). Three different factors affected the classification process; the first one is the makeup used in the photo (Figure 5-(a)), large-age difference (Figure 5-(b)), and color (Figure 5-(c)). We believe that those three factors also affected other kinship relationship categories. That is why the sister kinship relationship type achieved the highest average accuracy compared with other types of kinship.

To further improve the classifier's performance, we also conducted the testing process using four ensemble configurations, as shown in Table 2. As shown in Table 2, the ensemble configuration can improve the classifier's performance by around 0.5-0.8% compared with the single classifier configuration. As predicted, the second generation kinship relationship categories still produce the lowest average accuracy. Still, it is relatively higher compared with the single classifier results except for the grand mother-grand son kinship category. The best average accuracy of the ensemble classifier is 68.73% using Ensemble All (four of the DFaceNet classifiers).

Figure 6 shows the ROC curve plot of eleven kinship relationship categories from three different classifiers, the DFaceNet-FC512 classifier, DFaceNet-FC1K-CAtt classifier, and Ensemble All configuration. The AUC score is also included in the graph to provided insight information regarding the classifier. As shown in Figure 6, the ensemble configuration provides around 0.01 increase on the AUC score. The ROC curve of second-generation kinship relationship categories is not smooth with a lot of jigsaw-like lines, especially on grand father-grand daughter kinship.

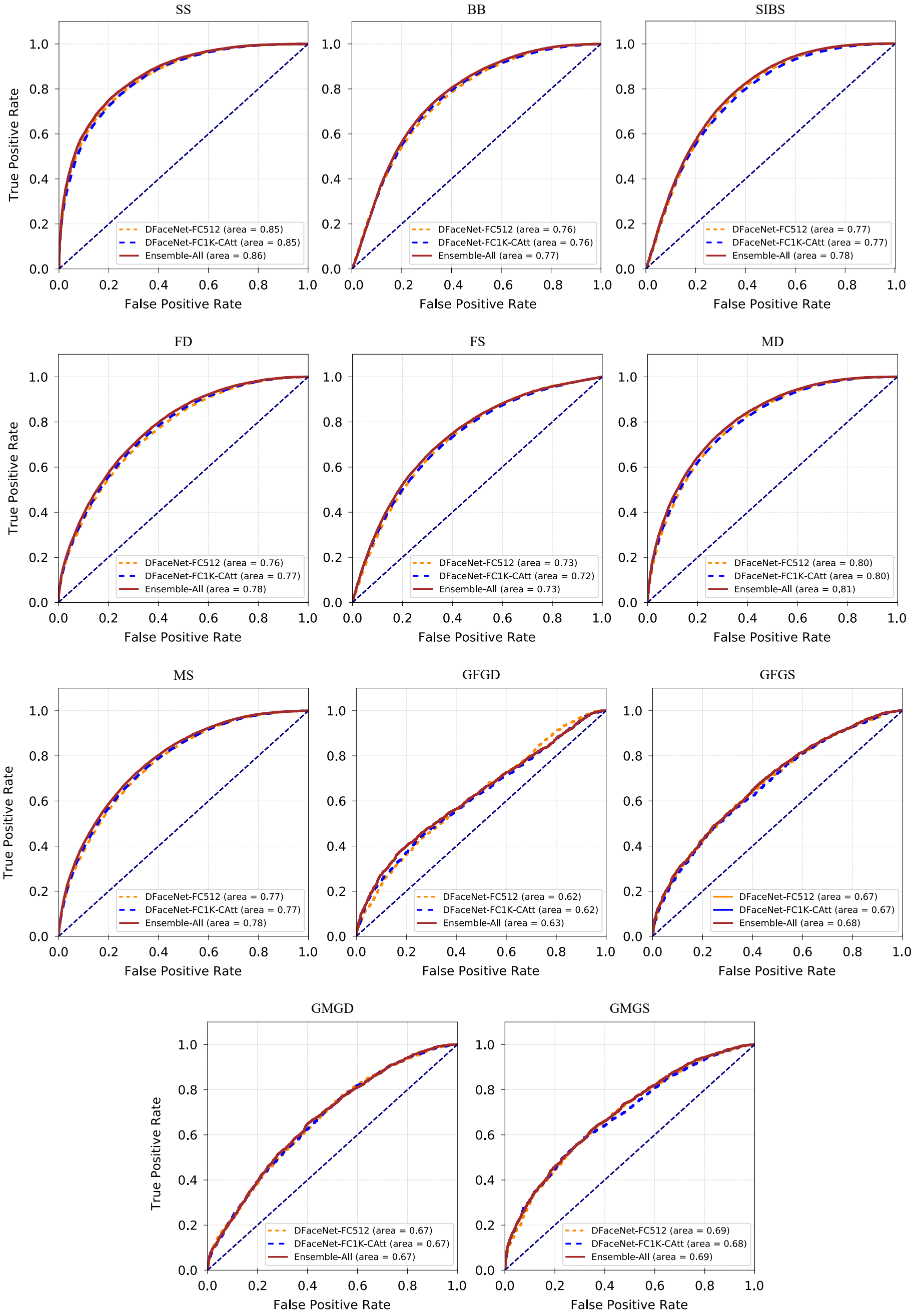


Figure 6: ROC curve of three different classifiers, the DFaceNet-FC512, DFaceNet-FC1K-CAtt, and Ensemble All, for 5-folds split configuration on FIW dataset.

Table 3: Comparison of our proposed classifier with several other methods on 5-folds FIW dataset.

No.	Method	Siblings	Parent	Grand	Avg.
			Child	Grand	All
1.	SphereFace [35]	73.15	69.76	65.60	69.18
2.	SDMLoss [49]	74.11	69.08	64.22	68.68
3.	DML[1] [50]	75.27	70.05	65.89	68.79
4.	Dual VGG-Face [34]	69.43	66.65	61.37	65.49
5.	FA-CNN [33]	73.64	71.12	62.93	68.84
6.	**Our method**	**72.68**	**70.94**	**63.53**	**68.73**

The AUC (area under the curve) on the ROC curve shows that the worst performance occurs in the grand father-grand daughter class and the best performance is occurs in the sister class.

Comparison with state-of-the-art (SOTA). We listed several different other methods that use 5-folds FIW dataset. In the early FIW dataset, the 5-folds configuration consists of only nine instead of eleven kinship categories [50]. Although Wang et al. [50] use different 5-folds configurations, we still included the results for information preservation because we cannot recreate the experiments due to no available information regarding the split configuration. Table 3 shows the comparison between our proposed classifier with several other methods on the 5-folds FIW dataset. We also include the average accuracy of each generation (siblings, parent-child, and grand parent-grand child) to provide more information regarding the classifier's performance on different generations.

4.3.3 RFIW'17

We use the RFIW2017 challenge split configurations to perform similar experiments as in the 5-folds experiments to make more comparisons. We use the same hyperparameter and epoch to perform the training process and tested using the validation dataset only because submission to the challenge website is already closed by the organizer. Table 4 shows the results of our proposed classifier on the RFIW'17

[1] The 5-folds dataset is different with nine kinship relationship categories instead of eleven.

Table 4: Verification results (%) on RFIW'17 validation dataset for Dual FaceNet with family-aware features and channel-based attention network.

#	Method	siblings			parent-child				Avg
		ss	bb	sibs	fd	fs	md	ms	
1.	DFaceNet-FC512-CAtt	75.5	73.4	72.5	67.5	67.7	70.8	70.3	71.16
2.	DFaceNet-FC1K-CAtt	77.9	71.6	71.9	67.6	67.9	70.6	70.6	71.20
3.	DFaceNet-FC2K-CAtt	76.0	72.1	71.6	67.6	68.0	70.6	69.8	70.86
4.	DFaceNet-FC512-CL-CAtt	76.4	70.9	70.9	66.7	67.3	69.9	68.6	70.15
5.	Ensemble 1 + 4	76.4	73.5	73.0	68.2	68.3	71.2	70.8	71.68
6.	Ensemble 2 + 4	**78.3**	72.6	72.2	68.3	68.5	70.8	70.7	71.66
7.	Ensemble 1 + 2 + 4	77.6	73.7	73.3	68.5	68.7	71.5	71.5	72.17
8.	Ensemble All	77.5	**73.9**	**73.5**	**69.1**	**69.2**	**71.8**	**71.7**	**72.44**

Table 5: Verification results (%) on RFIW'18 validation dataset for Dual FaceNet with family-aware features and channel-based attention network.

#	Method	siblings			parent-child				grandparent-grandchild				Avg
		ss	bb	sibs	fd	fs	md	ms	gfgd	gfgs	gmgd	gmgs	
1.	DFaceNet-FC512-CAtt	72.3	75.7	76.3	68.8	67.9	70.4	71.1	55.2	62.9	59.0	59.3	67.22
2.	DFaceNet-FC1K-CAtt	73.2	76.2	76.6	68.8	68.5	70.4	71.1	53.4	64.4	58.7	59.8	67,42
3.	DFaceNet-FC2K-CAtt	72.8	**76.5**	77.4	69.3	68.8	70.7	71.6	54.9	**65.5**	57.6	59.2	67.69
4.	DFaceNet-FC512-CL-CAtt	72.9	76.0	76.9	68.6	68.2	70.2	71.0	**56.0**	63.5	58.5	59.2	67.41
5.	Ensemble 1 + 4	73.2	76.2	77.5	69.4	68.5	71.1	71.8	55.6	63.7	57.9	59.8	67.73
6.	Ensemble 2 + 4	**73.6**	76.3	77.4	69.3	68.9	71.0	71.7	54.9	64.4	**59.1**	59.4	67.84
7.	Ensemble 1 + 2 + 4	73.6	76.3	77.4	69.7	68.9	71.3	72.0	54.8	64.7	58.6	**59.8**	67.96
8.	Ensemble All	73.5	76.4	**77.8**	**69.8**	**69.1**	**71.4**	**72.2**	55.3	64.4	58.3	**59.8**	**68.05**

dataset. As shown in Table 4, the best performance of single classifier configuration is achieved using the DFaceNet-FC1K-CAtt classifier with an average accuracy of 71.20%. The ensemble configuration is improved by around 0.5-1.0%, and the best performance is achieved using Ensemble All classifier with an average accuracy of 72.44%.

Same with previous experiments, we also plot the ROC curve of each kinship relationship category. Figure 7 shows the ROC curve of three different classifier configurations, including DFaceNet-FC512-CAtt, DFaceNet-FC1K-CAtt, Ensemble-1-2-4, and Ensemble All. As shown in Figure 7, the ROC analysis shows that our proposed classifier performs well with an AUC score of more than 80% except for the father-son and father-daughter kinship relationship. The same AUC score improvement of 0.01 as in the 5-folds experiments also occurs in the RFIW'17 experiments.

Table 6 shows the comparison of our proposed classifier with other methods on the RFIW'17 dataset. Unfortunately, we can also provide the accuracy on the validation set instead of the testing set because the organizer already close the submission server, and we don't have any annotation on the testing set. As shown in Table 6, our proposed classifier is comparable with other methods. We are aware that our proposed classifier does not produce the highest accuracy. Still, in

Table 6: Comparison of our proposed classifier with several other methods on RFIW'17 dataset (average accuracy of each category.

No.	**Method**	**Siblings**	**Parent Child**	**Avg. All**
1.	KinNet [21]	75.07	74.68	74.85
2.	AdvNet [11]	73.00	68.46	70.41
3.	LPQ-SIEDA [18]	54.53	55.01	54.81
4.	Multi-Set Learning [7]	63.68	62.66	63.10
5.	Parallel SPCNN [32]	62.01	60.81	61.33
6.	FA-CNN [33]	74.52	70.79	72.39
7.	**Our method**[2]	75.02	70.50	72.44

[2] The average accuracy is based on validation set instead of testing set

our understanding, the KinNet approaches [21] use a deeper and bigger classifier, which is natural will have more accuracy than our approaches.

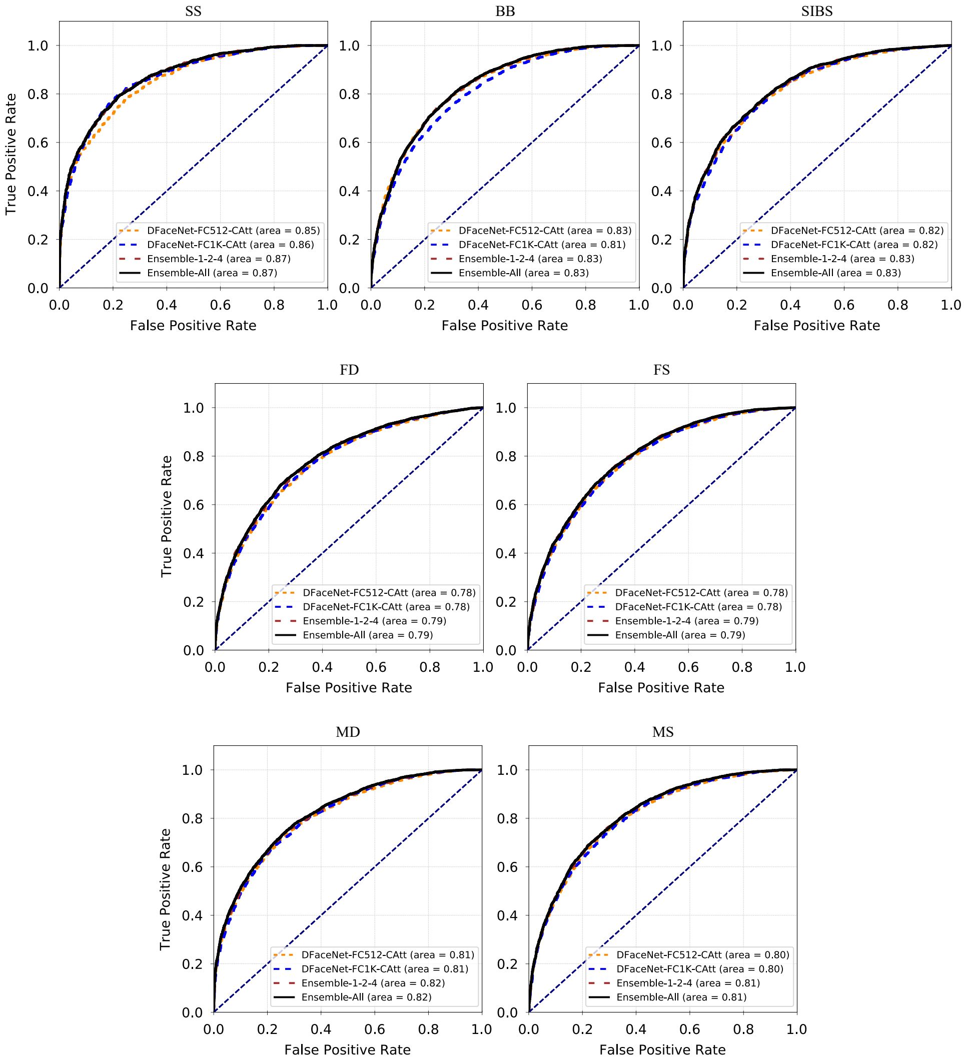


Figure 7: ROC curve of three different classifiers, the DFaceNet-FC512-CAtt, DFaceNet-FC1K-CAtt, Ensemble-1-2-4, and Ensemble All, for the RFIW'17 dataset.

4.3.4 RFIW'18

The last experiment is conducted using the RFIW'18 dataset. The RFIW'18 dataset is a subset of the FIW dataset used for the RFIW challenge 2018 and consists of the same number of kinship relationships as the 5-folds configuration. Same as the previous experiments, the same hyperparameter values were used to perform the experiments. Table 5 shows the results of the experiments using four different single classifiers along with four ensemble configurations. As shown in Table 5, the best single classifier performance is achieved using DFaceNet-FC2K-CAtt with an average accuracy of 67.69%. By using ensemble configuration, the classifier's performance is slightly improved by around 0.5%, with the best average accuracy of 68.05%. As we expected, the worst performance of the proposed classifier is on second generation relationship categories which also occurs in the previous experiments. The difference between RFIW'18 with two previous experiments is that the best performance is not occurring in the sister kinship category but in siblings kinship. We believe that those phenomena occur because the dataset's face images composition may consist of more face pairs with large-gap age.

Figure 8 shows the ROC curve of three different proposed classifiers, including DFaceNet-FC512-CL-CAtt, DFaceNet-FC2K-CAtt, and Ensemble-All. As shown in Figure 8, all classifier configurations do not perform well on the grand father-grand daughter and grand mother-grand daughter category. Same as in the 5-folds experiments, the best performance occurs in the same generation kinship relationships. According to Figure 8, the ensemble configuration can improve the AUC score by around 0.01 on all kinship relationship categories.

Table 7 shows the comparison of our proposed classifier with other methods

Table 7: Comparison of our proposed classifier with several other methods on RFIW'18 dataset (average accuracy of each category.

No.	**Method**	**Siblings**	**Parent**	**Grand**	**Avg.**
			Child	**Grand**	**All**
1.	Group #1 [6]	71.67	70.61	63.17	68.20
2.	Group #2	67.53	62.82	58.38	62.44
3.	Group #3	66.75	62.65	58.87	62.40
5.	FA-CNN [33]	70.34	68.54	62.83	66.96
6.	**Our method**	**70.71**	**69.51**	**61.62**	**66.97**

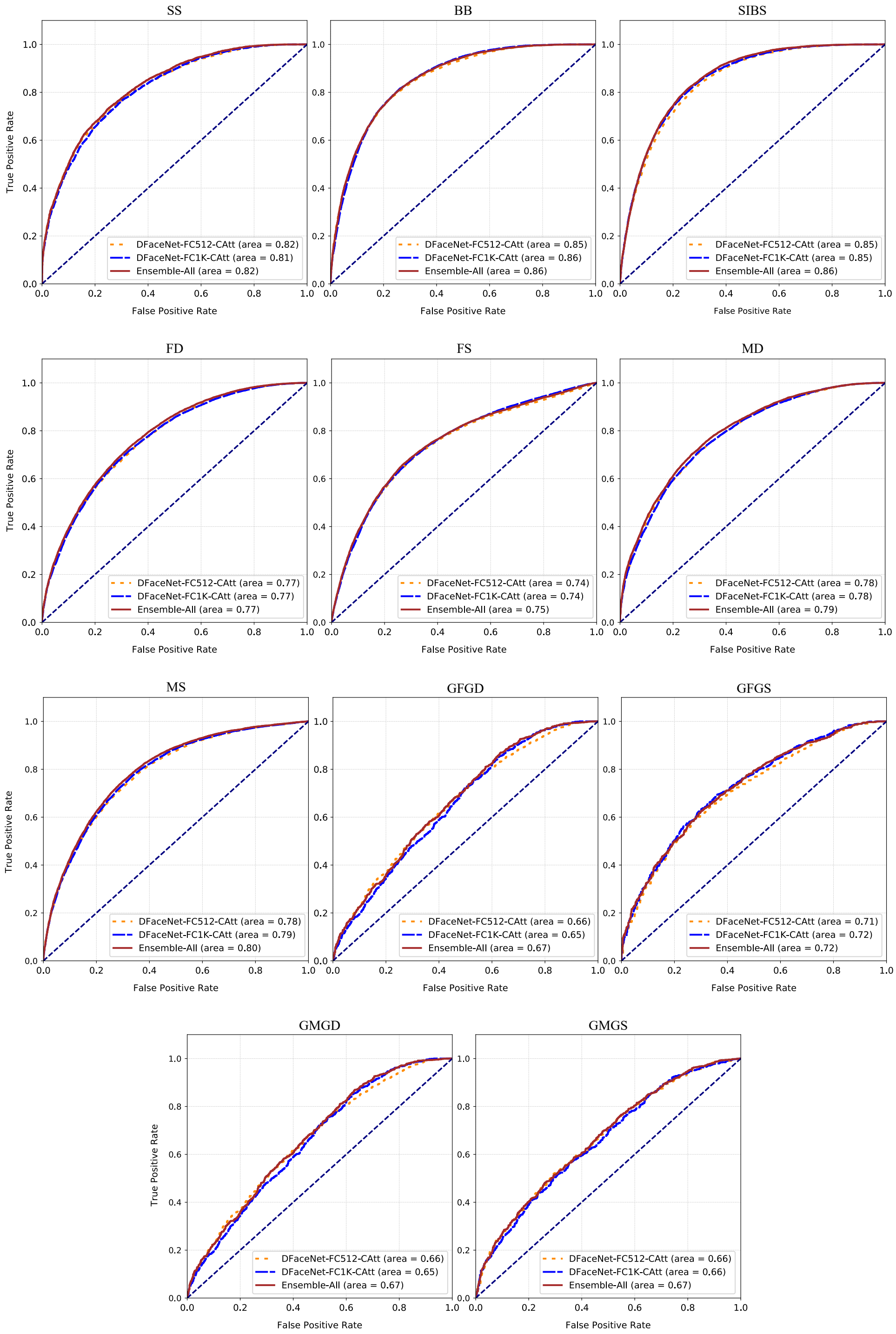


Figure 8: ROC curve of three different classifiers, the DFaceNet-FC512-CL-CAtt, DFaceNet-FC2K-CAtt, and Ensemble All, for RFIW'18 dataset.

on the RFIW'18 dataset. We took the three top participants with the highest performance on the RFIW'18 competition. Unfortunately, the method used by Group #2 and #3 is not published yet. As shown in Table 7, our proposed classifier can achieve an average accuracy of 66.97% and ranked the second-highest performance on the RFIW'18 dataset. Compared with the FA-CNN classifier, the proposed classifier produces a similar performance. Still, the performance per generation shows that the pyramid attention network can improve the performance on same and first-generation kinship relationships while decreasing the performance on second-generation kinship relationships.

5 Conclusion

We present our proposed classifier that combined FaceNet CNN architecture originally used for face recognition with pyramid attention network to solve the kinship verification problem. Our proposed classifier was formed by parallelling the FaceNet CNN architecture and adding family-aware features and a pyramid attention network. The final features were constructed by combining pyramid attention features and family-aware features and fed the features into three fully connected layers to perform the verification tasks. Experiments on three different subsets of the FIW dataset show that the proposed classifier can achieve good accuracy and is comparable with the state-of-the-art classifier on the FIW dataset. The proposed classifier achieves an average accuracy of 69.73% on the 5-folds RFIW dataset, 72.44% on the RFIW'17 dataset, and 66.97% on the RFIW'18 dataset.

For further study, experiments using several different CNN architectures (including non-face recognition architecture) with pyramid attention networks are demanding to show which CNN architectures perform best for image-based kinship verification. The second-generation kinship type may need to be solitary experimented due to lower facial features matched between the pair. Other concerns worth analyzing are the relation between each region of the face for kinship verification problems (e.g., eyes, lips, nose, etc.).

References

[1] Ambartsoumian, Artaches and Popowich, Fred. Self-Attention: A better building block for sentiment analysis neural network classifiers. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 130–139, 2018. DOI: 10.48550/arXiv.1812.07860.

[2] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015. DOI: 10.48550/arXiv.1409.0473.

[3] Britz, Denny, Goldie, Anna, Luong, Minh-Thang, and Le, Quoc. Massive exploration of neural machine translation architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1451, 2017. DOI: 10.48550/arXiv.1703.03906.

[4] Chen, Cunjian and Ross, Arun. Matching thermal to visible face images using a semantic-guided generative adversarial network. In *IEEE International Conference on Automatic Face & Gesture Recognition*, 2019. DOI: 10.1109/FG.2019.8756527.

[5] Choi, Yunjey, Choi, Minje, Kim, Munyoung, Ha, Jung-Woo, Kim, Sunghun, and Choo, Jaegul. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. DOI: 10.1109/CVPR.2018.00916.

[6] Dahan, Eran and Keller, Yosi. SelfKin: Self adjusted deep model for kinship verification. arXiv Preprint, 2018. DOI: 10.48550/arXiv.1809.08493.

[7] Dahan, Eran, Keller, Yosi, and Mahpod, Shahar. Kin-verification model on FIW dataset using multi-set learning and local features. In *Proceedings of the 2017 Workshop on Recognizing Families In the Wild*, RFIW '17, pages 31–35, New York, NY, USA, 2017. ACM. DOI: 10.1145/3134421.3134423.

[8] Dawson, Mitchell, Zisserman, Andrew, and Nellåker, Christoffer. From same photo: Cheating on visual kinship challenges. In *Asian Conference on Computer Vision*, pages 654–668. Springer, 2018. DOI: 10.1007/978-3-030-20893-6_41.

[9] Deng, Jiankang, Guo, Jia, Xue, Niannan, and Zafeiriou, Stefanos. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. DOI: 10.1109/CVPR.2019.00482.

[10] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. DOI: 10.48550/arXiv.1810.04805.

[11] Duan, Qingyan and Zhang, Lei. AdvNet: Adversarial contrastive residual net for 1 million kinship recognition. In *Proceedings of the 2017 Workshop on Recognizing Families In the Wild*, RFIW '17, pages 21–29, New York, NY, USA, 2017. ACM. DOI: 10.1145/3134421.3134422.

[12] Fang, R., Tang, K. D., Snavely, N., and Chen, T. Towards computational models of kinship verification. In *IEEE International Conference on Image Processing*, pages 1577–1580, 2010. DOI: 10.1109/ICIP.2010.5652590.

[13] Ge, Weifeng, Huang, Weilin, Dong, Dengke, and Scott, Matthew R. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–285, 2018. DOI: 10.1007/978-3-030-01231-1_17.

[14] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015. DOI: 10.1109/ICCV.2015.123.

[15] Jetley, Saumya, Lord, Nicholas A., Lee, Namhoon, and Torr, Philip H. S. Learn to pay attention. arXiv Preprint, 2018. DOI: 10.48550/ARXIV.1804.02391.

[16] Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, Jonathan, Girshick, Ross, Guadarrama, Sergio, and Darrell, Trevor. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 675–678, 2014. DOI: 10.1145/2647868.2654889.

[17] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[18] Laiadi, Oualid, Ouamane, Abdelmalik, Benakcha, Abdelhamid, and Taleb-Ahmed, Abdelmalik. RFIW 2017: LPQ-SIEDA for large scale kinship verification. In *Proceedings of the 2017 Workshop on Recognizing Families In the Wild*, RFIW '17, pages 37–39, New York, NY, USA, 2017. ACM. DOI: 10.1145/3134421.3134426.

[19] Laiadi, Oualid, Ouamane, Abdelmalik, Benakcha, Abdelhamid, Taleb-Ahmed, Abdelmalik, and Hadid, Abdenour. Multi-view deep features for robust facial kinship verification. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 877–881. IEEE, 2020. DOI: 10.1109/FG47880.2020.00118.

[20] Li, Lei, Feng, Xiaoyi, Wu, Xiaoting, Xia, Zhaoqiang, and Hadid, Abdenour. Kinship verification from faces via similarity metric based convolutional neural network. In *International Conference on Image Analysis and Recognition*, pages 539–548. Springer, 2016. DOI: 10.1007/978-3-319-41501-7_60.

[21] Li, Yong, Zeng, Jiabei, Zhang, Jie, Dai, Anbo, Kan, Meina, Shan, Shiguang, and Chen, Xilin. KinNet: Fine-to-coarse deep metric learning for kinship verification. In *Proceedings of the 2017 Workshop on Recognizing Families In the Wild*, RFIW '17, pages 13–20, New York, NY, USA, 2017. ACM. DOI: 10.1145/3134421.3134425.

[22] Liu, Jie, Zhang, Wenjie, Tang, Yuting, Tang, Jie, and Wu, Gangshan. Residual feature aggregation network for image super-resolution. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2359–2368, 2020. DOI: 10.1109/CVPR42600.2020.00243.

[23] Liu, Sifei, Yang, Jimei, Huang, Chang, and Yang, Ming-Hsuan. Multi-objective convolutional learning for face labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3451–3459, 2015. DOI: 10.1109/CVPR.2015.7298967.

[24] Liu, Weiyang, Wen, Yandong, Yu, Zhiding, Li, Ming, Raj, Bhiksha, and Song, Le. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017. DOI: 10.1109/CVPR.2017.713.

[25] Liu, Weiyang, Wen, Yandong, Yu, Zhiding, and Yang, Meng. Large-margin softmax loss for convolutional neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 507–516, 2016.

[26] Lu, Jiwen, Hu, Junlin, Zhou, Xiuzhuang, Shang, Yuanyuan, Tan, Yap-Peng, and Wang, Gang. Neighborhood repulsed metric learning for kinship verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2594–2601. IEEE, 2012. DOI: 10.1109/CVPR.2012.6247978.

[27] Lu, Jiwen, Zhou, Xiuzhuang, Tan, Yap-Pen, Shang, Yuanyuan, and Zhou, Jie. Neighborhood repulsed metric learning for kinship verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):331–345, 2013. DOI: 10.1109/TPAMI.2013.134.

[28] Lu, Xiaoqiang, Sun, Hao, and Zheng, Xiangtao. A feature aggregation convolutional neural network for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):7894–7906, 2019. DOI: 10.1109/TGRS.2019.2917161.

[29] Luong, Minh-Thang, Pham, Hieu, and Manning, Christopher D. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015. DOI: 10.48550/arXiv.1508.04025.

[30] Ma, Jiayi, Jiang, Xingyu, Fan, Aoxiang, Jiang, Junjun, and Yan, Junchi. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021. DOI: 10.1007/s11263-020-01359-2.

[31] Qin, X., Tan, X., and Chen, S. Tri-subject kinship verification: Understanding the core of a family. *IEEE Transactions on Multimedia*, 17(10):1855–1867, 2015. DOI: 10.1109/TMM.2015.2461462.

[32] Rachmadi, Reza Fuad and Purnama, I Ketut Eddy. Paralel spatial pyramid convolutional neural network untuk verifikasi kekerabatan berbasis citra wajah.

Jurnal Teknologi dan Sistem Komputer, 6(4):152–157, 2018. DOI: 10.14710/jtsiskom.6.4.2018.152-157.

[33] Rachmadi, Reza Fuad, Purnama, I Ketut Eddy, Nugroho, Supeno Mardi Susiki, and Suprapto, Yoyon Kusnendar. Family-aware convolutional neural network for image-based kinship verification. *International Journal of Intelligent Engineering and Systems*, 13(6):20–30, 2020. DOI: 10.22266/ijies2020.1231.03.

[34] Rachmadi, Reza Fuad, Purnama, I Ketut Eddy, Nugroho, Supeno Mardi Susiki, and Suprapto, Yoyon Kusnendar. Image-based kinship verification using dual VGG-Face classifier. In *IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pages 123–128. IEEE, 2021. DOI: 10.1109/IoTaIS50849.2021.9359720.

[35] Robinson, J. P., Shao, M., Wu, Y., Liu, H., Gillis, T., and Fu, Y. Visual kinship recognition of families in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2624–2637, 2018. DOI: 10.1109/TPAMI.2018.2826549.

[36] Robinson, Joseph P, Shao, Ming, and Fu, Yun. 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, 2021. DOI: 10.1109/TPAMI.2021.3063078.

[37] Robinson, Joseph P., Shao, Ming, Wu, Yue, and Fu, Yun. Families in the wild (FIW): Large-scale kinship image database and benchmarks. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 242–246, New York, NY, USA, 2016. ACM. DOI: 10.1145/2964284.2967219.

[38] Robinson, Joseph P, Shao, Ming, Zhao, Handong, Wu, Yue, Gillis, Timothy, and Fu, Yun. Recognizing families in the wild (RFIW) data challenge workshop in conjunction with ACM MM 2017. In *Proceedings of the 2017 Workshop on Recognizing Families in the Wild*, pages 5–12, 2017. DOI: 10.1145/3134421.3134424.

[39] Robinson, Joseph P., Yin, Yu, Khan, Zaid, Shao, Ming, Xia, Siyu, Stopa, Michael, Timoner, Samson, Turk, Matthew A., Chellappa, Rama, and Fu, Yun. Recognizing families in the wild (RFIW): The 4th edition. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 857–862, 2020. DOI: 10.1109/FG47880.2020.00138.

[40] Schroff, Florian, Kalenichenko, Dmitry, and Philbin, James. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. DOI: 10.1109/CVPR.2015.7298682.

[41] Sukhbaatar, Sainbayar, Weston, Jason, Fergus, Rob, et al. End-to-end memory networks. *Advances in Neural Information Processing Systems*, 28:2440–2448, 2015.

[42] Sun, Yi, Chen, Yuheng, Wang, Xiaogang, and Tang, Xiaoou. Deep learning face representation by joint identification-verification. *Advances in Neural Information Processing Systems*, 27:1988–1996, 2014.

[43] Sun, Yi, Wang, Xiaogang, and Tang, Xiaoou. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014. DOI: 10.1109/CVPR.2014.244.

[44] Tang, Gongbo, Müller, Mathias, Gonzales, Annette Rios, and Sennrich, Rico. Why self-attention? A targeted evaluation of neural machine translation architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4263–4272, 2018. DOI: 10.48550/arXiv.1808.08946.

[45] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[46] Wang, Fei, Jiang, Mengqing, Qian, Chen, Yang, Shuo, Li, Cheng, Zhang, Honggang, Wang, Xiaogang, and Tang, Xiaoou. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. DOI: 10.1109/CVPR.2017.683.

[47] Wang, Feng, Xiang, Xiang, Cheng, Jian, and Yuille, Alan Loddon. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 1041–1049. ACM, 2017. DOI: 10.1145/3123266.3123359.

[48] Wang, Hao, Wang, Yitong, Zhou, Zheng, Ji, Xing, Gong, Dihong, Zhou, Jingchao, Li, Zhifeng, and Liu, Wei. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. DOI: 10.1109/CVPR.2018.00552.

[49] Wang, S., Ding, Z., and Fu, Y. Cross-generation kinship verification with sparse discriminative metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018. DOI: 10.1109/TPAMI.2018.2861871.

[50] Wang, Shuyang, Robinson, Joseph P, and Fu, Yun. Kinship verification on families in the wild with marginalized denoising metric learning. In *12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 216–221, 2017. DOI: 10.1109/FG.2017.35.

[51] Wen, Yandong, Zhang, Kaipeng, Li, Zhifeng, and Qiao, Yu. A discriminative feature learning approach for deep face recognition. In *PRoceedings of the European Conference on Computer Vision*, pages 499–515. Springer, 2016. DOI: 10.1007/978-3-319-46478-7_31.

[52] Woo, Sanghyun, Park, Jongchan, Lee, Joon-Young, and So Kweon, In. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. DOI: 10.1007/978-3-030-01234-2_1.

[53] Wu, Xiaoting, Feng, Xiaoyi, Cao, Xiaochun, Xu, Xin, Hu, Dewen, López, Miguel Bordallo, and Liu, Li. Facial kinship verification: A comprehensive review and outlook. *International Journal of Computer Vision*, pages 1–32, 2022. DOI: 10.1007/s11263-022-01605-9.

[54] Yan, Haibin and Hu, Junlin. Video-based kinship verification using distance metric learning. *Pattern Recognition*, 75:15–24, 2018. DOI: 10.1016/j.patcog.2017.03.001.

[55] Yang, Zichao, Yang, Diyi, Dyer, Chris, He, Xiaodong, Smola, Alex, and Hovy, Eduard. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

[56] Yu, Jun, Li, Mengyan, Hao, Xinlong, and Xie, Guochen. Deep fusion siamese network for automatic kinship verification. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 892–899. IEEE, 2020. DOI: 10.1109/FG47880.2020.00127.

[57] Zhang, Han, Goodfellow, Ian, Metaxas, Dimitris, and Odena, Augustus. Self-attention generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, pages 7354–7363, 2019. DOI: 10.48550/arXiv.1805.08318.

[58] Zhao, Ting and Wu, Xiangqian. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3085–3094, 2019. DOI: 10.1109/CVPR.2019.00320.

Received 12th March 2022