Kinship Verification from Ear Images: An Explorative Study with Deep Learning Models

Grega Dvoršak[1], Ankita Dwivedi[2], Vitomir Štruc[3], Peter Peer[1], Žiga Emeršič[1]
[1]*University of Ljubljana, Faculty of Computer and Information Science, Vecna pot 113, SI-1000 Ljubljana, Slovenia*
[2] *Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh 226031, India*
[3]*University of Ljubljana, Faculty of Electrical Engineering, Trzaska 25, SI-1000 Ljubljana, Slovenia*
E-mail: gdvorsak97@gmail.com

***Abstract*—The analysis of kin relations from visual data represents a challenging research problem with important real-world applications. However, research in this area has mostly been limited to the analysis of facial images, despite the potential of other physical (human) characteristics for this task. In this paper, we therefore study the problem of kinship verification from ear images and investigate whether salient appearance characteristics, useful for this task, can be extracted from ear data. To facilitate the study, we introduce a novel dataset, called KinEar, that contains data from 19 families with each family member having from 15 to 31 ear images. Using the KinEar data, we conduct experiments using a Siamese training setup and 5 recent deep learning backbones. The results of our experiments suggests that ear images represent a viable alternative to other modalities for kinship verification, as 4 out of 5 considered models reach a performance of over 60% in terms of the Area Under the Receiver Operating Characteristics (ROC-AUC).**

***Index Terms*—biometrics, kinship, ear biometrics, computer vision**

I. Introduction

Kinship verification is an important computer vision problem, where the goal is to determine whether two people are in a kin relation or not based on the analysis of visual correspondences. It has important applications in various areas, ranging from locating missing children to social media analyses [26], [32]. Most of the research on this topic has so far been done using face images[1], where diverse datasets with large numbers of families and family members are publicly available [26]. Research with other modalities, on the other hand, is still largely missing from the literature. We therefore address this gap in this paper and study the potential of ear images for kinship verification, as illustrated in Fig. 1.

While ear recognition gained on popularity in recent years within the biometric community [8], [10], [29], the analysis of kin relations based on ear images is a relatively unexplored problem. To the best of our knowledge, there is only a single work available in the literature that studies kinship verification from ear images using a model-based approach with hand-crafted features [20]. Here, the authors report promising initial results on a dataset of 134 images with 21 families (i.e., 21 parents and 25 children). While the aforementioned study explored whether geometric and low-level structural characteristics are shared between first-degree blood relatives,

[1]Or multi-modal solutions involving face images [33].

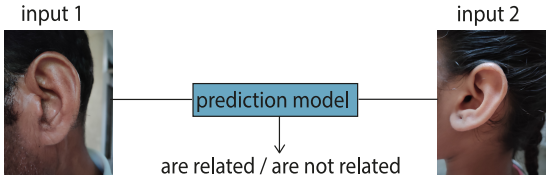

Fig. 1: **Illustration of the kinship verification task.** In this paper, we investigate the problem of kinship verification from ear images using automatically learned appearance features extracted by deep learning models.

we take a step further in this paper and investigate whether higher-level appearance cues can be exploited for kinship verification from ear images. To this end, we analyze five recent deep learning models, including popular convolutional neural networks (CNNs), attention-augmented models and transformers, which were shown in the literature to be able to extract highly descriptive image representations with rich semantic information for various vision tasks. As part of our research, we aim to answer the following research questions: (i) Is kinship verification from ear images feasible using contemporary appearance-based representations? (ii) What performance can be expected in this setting? (iii) How does the observed performance correlate to differences in the data representation? (iv) What are the limitations of appearance-based kinship verification that cause errors in the predictions?

To help answer the above questions and facilitate the study, we introduce a novel dataset of ear images with annotated kin relations. The dataset, named KinEar, is the first publicly available dataset in this problem domain and represents one of the major contributions of this work. KinEar contains data on 19 families with 76 subjects and consists of 1477 images that correspond to a total of $37,282$ kin pairs that can be used to train, test and analyze the performance of kinship verification models. The introduced dataset is, hence, more than an order of magnitude larger than the dataset from [20] in terms of images as well as kin pairs.

In summary, the main contributions of this paper can be grouped into the following main points:

- **Feasibility study:** We are the first to study kinship verification from ear images with powerful appearance features extracted with contemporary deep learning models.

- **The KinEar dataset:**[2] We introduce a novel dataset of ear images suitable for studying (visual) kinship recognition models and make it publicy available to the research community through: http://ears.fri.uni-lj.si/.
- **Comprehensive analysis:** We report important findings with respect to the ear-based kinship verification task across different deep learning models and qualitatively analyze failure as well as success (verification) cases that provide insight into the feasiblity of the task with image representations extracted with deep learning models.

II. Related Work

In this section, we position our work within the existing literature. Specifically, we discuss existing work on (i) visual kinship recognition, (ii) ear recognition in the context of biometrics, and (iii) kinship recognition from ear images.

A. Visual Kinship Recognition

Existing work on kinship recognition from visual data is predominantly focused on analyzing facial images, mostly due to the availability of suitable datasets, which are scarce for other modalities. One such work by Lu *et al.* [19], for example, presented a Discriminative Deep Metric Learning (DDML) method for face kinship verification with images captured in-the-wild. In DDML, a discriminative neural network is used to project face pairs into the same latent feature space, learned by minimizing distances between positive image pairs and maximizing distances between negative pairs. Wu *et al.* [34] proposed Latent Adaptive Subspace (LAS) learning for kinship classification by using an auxiliary dataset to address the problem of unavailable children data for training. Nandy *et al.* [22] presented a deep learning approach using a Siamese convolutional neural network architecture to quantify the similarity between two images and solve the kinship verification task on the Families in the Wild dataset [26]. A large amount of work along these lines has been presented in the literature over the years, as evidenced by recent surveys on this topic [24], [25]. The majority of this work explores metric-learning solutions and Siamese model topologies that allow to derive a measure of kinship using a pair of input images. We follow these trends and also investigate a Siamese model setup as the basis for ear-based kinship verification.

B. Ear Recognition

Ear recognition techniques have evolved from early approaches based on geometric, structural and low-level texture features [4], [15], [21] to more recent deep learning solutions [7], [9]. This development led to significant performance improvements that now allow to deploy ear recognition models across images captured in unconstrained settings [10], [11].

Within these developments, Sinha *et al.* [28] proposed a solution that relied on histograms of oriented gradients (HoGs) with support vector machines (SVMs) for ear localization before using a deep neural network for recognition. Stepec *et al.* [29] presented a deep constellation model for ear recognition that used global as well as local ear characteristics to generate descriptive representations for recognition purposes. Alshazly *et al.* [2] performed a study on ear recognition with various convolutional neural networks (CNNs) such as AlexNet [17], VGG [23], Inception [30], ResNet [13] and ResNeXt [35] and reported highly competitive performance on the unconstrained EarVN 1.0 dataset [14]. Emersic *et al.* [6] introduced a complete ear recognition pipeline based on convolutional neural networks. Alshazly *et al.* [1] investigated a system for ear recognition based on ensembles of CNN models. Different networks of increasing depth were trained in this work and the best models were used to build ensembles to improve performance.

Inspired by the success of deep learning models for ear recognition and their ability to learn powerful data representations by correlating salient image characteristic to the provided reference labels, we explore in this paper the possibility of automatically learning high-level ear representations for kinship verification.

C. Kinship Recognition Using Ear Images

Kinship recognition from ear images is a relatively unexplored field with only a single paper by Meng *et al.* [20] available in the literature. As already outlined in the introductory section, the authors propose a model-based approach for kinship and gender verification using hand-crafted features. The approach first performs viewpoint correction and then generates ear descriptions using 81 triangles with 9 common points on every ear image. The generated features are then used to determine kinship from ear-image pairs. While we address the same conceptual tasks as the work in [20], the recognition approach used in this paper differs in several aspects from the research of Meng *et al.*, i.e.: (i) we automatically learn features for kinship verification instead of using hand-crafted representations, (ii) our approach is model-free and relies on the analysis of raw appearance information, (iii) in addition to the data representation we also learn a classifier that is directly applicable for the kinship verification tasks, and (iv) we conduct experiments over a (novel) larger dataset with a larger number of subjects and image pairs.

III. Methodology

In this section we present the methodology of our work. We start the section with a formal problem formulation, then describe the basic framework utilized, the deep learning models considered, and finally introduce the novel KinEar dataset and the performance metrics used in the experimental evaluation.

A. Problem Formulation

Kinship verification from ear images represents a two-class problem where the goal is to determine if the ears in the two input images come from subjects in a kin relation or not, as already illustrated in Fig. 1. Let $x_1 \in \mathbb{R}^{w \times h \times 3}$ and $x_2 \in \mathbb{R}^{w \times h \times 3}$ represent two RGB ear images, and let ψ be a kinship

[2]The dataset will posted after the review procedure.

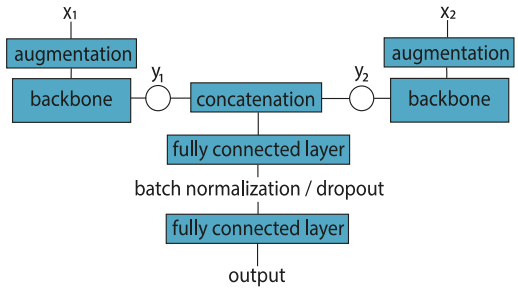

Fig. 2: **Overview of the utilized framework.** In accordance with existing literature in related problem areas, we use a Siamese model setup as the basis for the study, which can be conveniently implemented using various backbones.

verification model trained to produce a kinship score given x_1 and x_2 as input. The kinship verification task then assigns the input pair (x_1, x_2) to either the class of kin-related images w_1 or the class of images without a kin relation w_2, or formally:

$$(x_1, x_2) = \begin{cases} w_1, & \text{if } \psi(x_1, x_2) > \Delta \\ w_2, & \text{otherwise} \end{cases}, \tag{1}$$

where Δ represents a decision threshold that controls the trade-off between false positives and false negatives.

B. The Experimental Framework

Following the dominant approaches from the face-related kinship-recognition literature [24], [25], we use a Siamese framework as the basis for our study. A high-level overview of the adopted framework is presented in Fig. 2.

Architecture. The framework uses a Siamese model architecture that takes two ear images, x_1 and x_2, of different people as input. The two branches of the Siamese architecture share parameters and are implemented with a selected backbone model. These backbones produce image embeddings (or image representations), y_1 and y_2, that are concatenated and fed to a couple of fully connected layers that capture the correlations between the two embeddings and ultimately generate the kinship score $\psi(x_1, x_2)$ that determines whether the people in the input images are related or not.

Training. When training the Siamese model we use binary supervision for image pairs with and without kin relations and select binary cross-entropy as the learning objective. While the KinEar dataset (introduced later) offers a larger amount of data than previous datasets in this area, data augmentation is still needed to ensure that the models do not overfit. During the training procedure, we, therefore, utilize the functionality from the Keras and Torchvision libraries and use: (i) random translation in all directions by a factor of 0.1, (ii) random rotation in both geometric directions for an angle between 0 and 45 degrees, (iii) random zoom by a width and height factor of 0.1 and (iv) random horizontal flip. When using the transformer-based CoTNet backbone [18], we use additional data augmentation because transformer-based models typically require more data. Along with the previously mentioned functionalities, we use: (i) color jitter with brightness and hue set to 0.2, (ii) Gaussian blur with kernel size 5×9 and σ between 0.1 and 5, (iii) random solarization and (iv) random sharpness adjustment with the sharpness factor set to 2. Batch normalization and dropout layers are included in the model to improve training characteristics.

C. The Backbone Models

We use five different backbone models within the framework described above to extract image representations for ear-based kinship verification. The considered backbones are all publicly available to foster reproducibility and are selected due to their state-of-the-art performance for different vision tasks. Details on the backbones are given below.

- **VGG16:** The first backbone is the VGG16 model proposed initialy by Parkhi *et al.* [23]. As the name suggests, the model consist of 16 convolutional layers with small kernels (of size 3×3) and interspersed max-pooling layers. For the experiments, we use the model pretrained for face recognition on the VGG Face dataset.
- **ResNet-152:** The second backbone considered is a ResNet-152. This model comes from the family of ResNet models, developed by He *et al.* [13], that allow for efficient learning of very deep convolutional networks due to the presence of skip connections. The selected backbone, ResNet-152, contains 152 convolutional layers, but is lighter and less complex than VGG16. For the experiments, we again start with pretrained weights (from ImageNet [27]) to have a good initialization for training.
- **USTC-NELSLIP:** The third backbone used is a ResNet-50, which was shown to achieve the best overall performance within the USTC-NELSLIP model. Here, the USTC-NELSLIP model, initially presented by Yu *et al.* [36] for kinship verification using face images, is very similar in the overall design as our adopted framework. We, therefore select ResNet-50 as a light weight alternative to the ResNet-152 presented above. We again use existing weights pretrained on the VGG Face 2 dataset [3] to initialize the model for the experiments.
- **AFF:** The fourth backbone is the Attentional Feature Fusion (AFF) model, presented by Dai *et al.* [5], where the main idea is to implement feature fusion through an attention mechanism instead of simple operations like concatenation or addition. The main motivation behind using the AFF model in our experiments is to fuse the global and local information in ear images through an attention mechanism and then exploit the computed representations for kinship verification.
- **CoTNet:** The last backbone considered in our experiments is the Contextual Transformer Network, presented by Li *et al.* in [18]. CoTNet uses a transformer architecture to exploit contextual information with the attention mechanism. The goal of the model is to exploit the entire context available in the input data to improve the learning of the attention matrix and consequently improve performance across various computer vision tasks. This

Characteristic	Value
#Family	19
#Members/Subjects	76
#Images	1477
#Subject-2-Subject Kin Relations	96
#Kin image pairs	37282
Minimal Ear Resolution	250×250
Capture devices	Various

TABLE I: **Summary statistics for the KinEar dataset.** The number of kin relations provided is only for positive pairs, i.e., true kin relationships.

is achieved by the proposed CoT block which is an alternative to standard convolutions in CNNs.

D. The KinEar Dataset

To the best of our knowledge, no dedicated dataset is publicly available for research into kinship verification from ear images. We, therefore, introduce a new dataset for this task in this section and make it publicly available for research purposes from: http://ears.fri.uni-lj.si/.

The dataset, named KinEar, was acquired at the Dr. A.P.J. Abdul Kalam Technical University and in collaboration with members of the University of Ljubljana. The dataset contains data on 19 families with ear images for each family member. The total number of images in the dataset is 1477. There are 19 fathers, 19 mothers, 19 sons and 17 daughters present in the KinEar dataset. Manually annotated bounding boxes are provided for all ear images in order to ensure that all ears in the dataset can be properly aligned. While the capture devices and consequent image characteristics varied from family to family, we made sure that the ear region is always at least 250×250 pixels in size, which provides a reasonable resolution useful for most modern CNN-based models. Details on the dataset can be found in Table I and a per–family break down in Table II. Fig. 3 shows the distribution of relationships of family members over the entire dataset and Fig. 4 presents a few visual examples.

E. Experimental Setup

When training the models, pairs of images corresponding to different people are sampled from the dataset in a way that ensures that an approximately equal number of positive and negative pairs is included in each training batch. The ground truth is determined from a list provided with the dataset where related family member pairs are listed. The pairs of images that are not listed are considered to be negative examples in the training. The husband and wife of a family are regarded as not related. The training set contains 14 families, the validation set contains 2 and the testing set contains 3 families. The number of all possible image pairs in the testing set is 12960, out of which 9692 are negative and 3268 are positive.

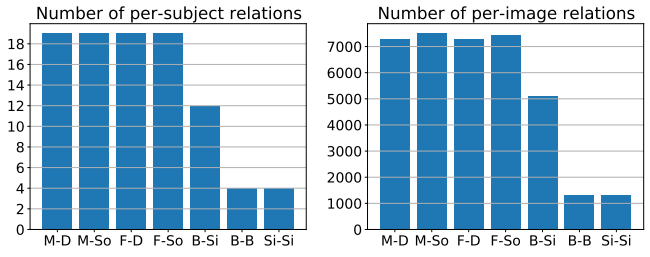

Fig. 3: **The distribution of family relations in the KinEar dataset**. The family members are represented as: F – father, M – mother, So – son, D – daughter, B – brother, Si – sister.

Family #	#Members	#Images	#Relations	#Kin pairs
1	4	74	5	1649
2	4	60	5	1125
3	4	84	5	2300
4	4	65	5	1339
5	4	60	5	1125
6	4	60	5	1125
7	4	75	5	1799
8	4	91	5	2599
9	4	96	5	2793
10	4	99	5	2903
11	4	87	5	2375
12	4	84	5	2183
13	4	83	5	2094
14	4	84	5	2133
15	5	103	9	3755
16	4	89	5	2484
17	3	45	2	450
18	4	64	5	1308
19	4	74	5	1743

TABLE II: **Per–family summary of the KinEar dataset.** The number of relations and image pairs is counted only for the positive pairs.

Fig. 4: **Examples of image pairs from the KinEar dataset**. The presented images are aligned and resized in accordance with the provided bounding boxes.

We use different performance indicators when testing the considered backbones within our overall framework. The performance indicators are: (i) classification accuracy (CA), (ii) sensitivity, (iii) specificity, and (iv) the area under the Receiver Operating Characteristics (ROC) curve (ROC-AUC). The first three are defined with the following equations [12]:

$$CA = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{3}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{4}$$

Backbone	CA [%]	Se. [%]	Sp. [%]	ROC-AUC [%]
VGG16 [23]	64.01	64.01	64.01	69.22
ResNet152 [13]	57.50	57.50	57.51	63.14
USTC-NELSLIP [36]	55.12	55.14	55.10	57.29
AFF [5]	60.00	60.00	60.00	64.01
CoTNet [18]	61.85	61.84	61.86	65.88

TABLE III: **Kinship verification performance for different backbones.** The results are reported for decision thresholds at the equal error operating point of the ROC curves. The "Se." column represents the sensitivity and the "Sp." column represents the specificity.

where FP represents the number of false positive classifications, FN the number of false negatives, TP the number of true positives, and TN the number of true negative classifications. Classification accuracy is an often used metric, which provides information on the ratio between correctly classified cases against all cases. Sensitivity conveys the ratio between the number of pairs classified as positive and the number of all pairs that are truly related, and specificity captures the ratio between the number of pairs classified as negative and all pairs that are truly not related.

The main performance metric that we use in our experiments is the ROC-AUC, which is the area under the ROC curve, a graph created by plotting the true positive rate against the false positive rate at different threshold values Δ [31], [37].

F. Experimental Details

The experiments were performed on an Nvidia GeForce GTX 1060 graphics card with 6 GB of video RAM. We use the Adam optimizer [16] with the learning rate equal to 10^{-5} for the VGG16, ResNet152 and USTC-NELSLIP models and 10^{-4} for the AFF and CoTNet models. During testing, the AFF and CoTNet models' thresholds can be set between -1 and 1, while the remaining models' thresholds can be set between 0 and 1.

IV. Experiments and Results

A. Quantitative Results

Table III shows the results of the performance evaluation with all considered backbone models at a decision threshold Δ_{eer} that ensures equal errors on the ROC curves in Fig. 5.

As can be seen, the best performing model is VGG16. Four out of the five models reach an ROC-AUC score of at least 60%. The CoTNet and AFF models are second and third best, which indicates that using transformer-based architecture and efficient feature fusion is beneficial for performance. Another observation of the results is that generally the models with lower depth are more successful. This is shown not only in the fact that VGG16, which has the fewest number of layers, achieved the best results, but also in the fact that when comparing backbones for the AFF and CoTNet models, where backbones of different depth can be chosen, the best results are obtained when using the backbone with the fewest layers possible. Given that VGG16 is not the most lightweight network despite its shallow design compared to the other

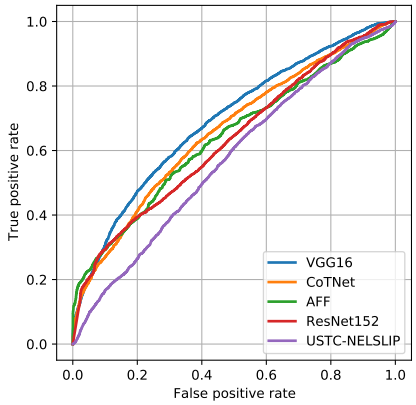

Fig. 5: **ROC curves of the comparative evaluation.** The VGG16 model exhibits the strongest performance, followed in order by CoTNet, AFF, ResNet-152, and USTC-NELSLIP.

models, we hypothesize that data representations with limited abstraction computed through a lower number of layers are more informative for the ear-based kinship verification task. However, additional research with in depth analysis of the characteristics of the learned representations are needed to validate this hypothesis.

B. Qualitative Results

In Fig. 6 we show some qualitative results obtained with the best performing VGG16 model at the decision threshold of Δ_{eer} with pairs of images that resulted in false positive, false negative, true positive and true negative results. Two examples of each result are shown. The left false negative pair shows the ears of a father and his daughter, meaning there is a difference in gender which makes it more difficult for the models to predict the correct result. There are some differences in the shape, particularly on the lower part of the ear, along with some difference in illumination conditions as well. The right false negative pair likely occurred due to one of the ears being captured at a significantly different angle, which makes the images appear more dissimilar. Among the false positives, the left pair example show ears which have similar shapes on their outer edges. Because both persons are children, it is harder to distinguish between them compared to the case when one person is a grown up and the other is a child. The right example

Fig. 6: **Sample qualitative results.** Pairs of input ear images are shown that resulted in false negative, false positive, true positive and true negative results with the VGG16 model at a decision threshold of Δ_{eer} that ensured equal errors on the corresponding ROC curve.

is similar, as the subjects are both daughters in their families. In general, there are few false positive cases that occur with all models. Both true negative pairs show some differences in the ear shapes and the difference in illumination conditions helps with the correct predictions as well. Lastly, the true positive pairs show ears which are visually similar. These images are taken under similar conditions such as illumination, angle and distance, and are therefore easy to classify correctly. Cases such as the true positive and true negative examples are among the easier ones to predict for all models. Overall, this analyies suggest the the external conditions have a considerable impact of kinship verification performance pointing to a need for controlled capture conditions or normalization techniques that can normalize the data prior to the verification process.

V. Conclusion

In this paper, we studied the task of visual kinship verification from ear images using a number of deep learning models utilized within a Siamese learning framework. Additionally, we also presented a new dataset, called KinEar, that contains ear images of members of 19 families and is made available to the research community. The results of our performance evaluation showed that ears are a suitable modality for kinship verification, especially when learned data representations are used for image description. In our setting, 4 out of the 5 considered models achieved an ROC-AUC score of at least 60%, the best model being VGG16 with an ROC-AUC score of 69.2%. Given that state-of-the-art solution based on facial images report performance only somewhat above these scores, kinship analysis from ear images certainly represents a topic worth investigating further.

References

[1] H. Alshazly, C. Linse, E. Barth, and T. Martinetz. Ensembles of deep learning models and transfer learning for ear recognition. In *Sensors*, volume 19, page 4139, 2019.

[2] H. Alshazly, C. Linse, E. Barth, and T. Martinetz. Deep convolutional neural networks for unconstrained ear recognition. In *IEEE Access*, volume 8, pages 170295–170310, 2020.

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018.

[4] M. Choras and R. S. Choras. Geometrical algorithms of ear contour shape representation and feature extraction. In *International Conference on Intelligent Systems Design and Applications*, pages 451–456, 2006.

[5] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3560–3569, 2021.

[6] Ž. Emeršič, J. Križaj, V. Štruc, and P. Peer. Deep ear recognition pipeline. In *Recent Advances in Computer Vision*, pages 333–362, 2019.

[7] Ž. Emeršič, B. Meden, P. Peer, and V. Štruc. Covariate analysis of descriptor-based ear recognition techniques. In *International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, pages 1–9, 2017.

[8] Ž. Emeršič, B. Meden, P. Peer, and V. Štruc. Evaluation and analysis of ear recognition models: performance, complexity and resource requirements. In *Neural Computing and Applications*, volume 32, 2020.

[9] Ž. Emeršič, N. Playa, V. Štruc, and P. Peer. Towards accessories-aware ear recognition. In *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–8, 2018.

[10] Ž. Emeršič, V. Štruc, and P. Peer. Ear recognition: More than a survey. In *Neurocomputing*, volume 255, pages 26–39, 2017.

[11] Ž. Emeršič, D. Sušanj, B. Meden, P. Peer, and V. Štruc. Contexednet: Context–aware ear detection in unconstrained settings. *IEEE Access*, 9:145175–145190, 2021.

[12] R. Gajšek, V. Štruc, F. Mihelič, A. Podlesek, L. Komidar, G. Sočan, and B. Bajec. Multi-modal emotional database: Avid. *Informatica*, 33(1), 2009.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] V. T. Hoang. Earvn1.0: A new large-scale ear images dataset in the wild. In *Data in brief*, volume 27. Elsevier, 2019.

[15] D. J. Hurley, M. S. Nixon, and J. N. Carter. Automatic ear recognition by force field transformations. In *IEE colloquium on visual biometrics*, pages 7–1, 2000.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ArXiv*, 2014.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.

[18] Y. Li, T. Yao, Y. Pan, and T. Mei. Contextual transformer networks for visual recognition. In *ArXiv*, 2021.

[19] J. Lu, J. Hu, and Y. P. Tan. Discriminative deep metric learning for face and kinship verification. In *IEEE Transactions on Image Processing*, volume 26, pages 4269–4282, 2017.

[20] D. Meng, M. S. Nixon, and S. Mahmoodi. Gender and kinship by model-based ear biometrics. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2019.

[21] B. Moreno, A. Sanchez, and J. F. Vélez. On the use of outer ear images for personal identification in security applications. In *Proceedings IEEE 33rd Annual 1999 International Carnahan Conference on Security Technology*, pages 469–476, 1999.

[22] A. Nandy and S. S. Mondal. Kinship verification using deep siamese convolutional neural network. In *14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–5, 2019.

[23] O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

[24] X. Qin, D. Liu, and D. Wang. A literature survey on kinship verification through facial images. *Neurocomputing*, 377:213–224, 2020.

[25] J. P. Robinson, M. Shao, and Y. Fu. Survey on the analysis and modeling of visual kinship: A decade in the making. *arXiv preprint arXiv:2006.16033*, 2020.

[26] J. P. Robinson, M. Shao, Y. Wu, H. Liu, T. Gillis, and Y. Fu. Visual kinship recognition of families in the wild. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2624–2637, 2018.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, J. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. In *International Journal of Computer Vision (IJCV)*, volume 115, pages 211–252, 2015.

[28] H. Sinha, R. Manekar, Y. Sinha, and P. K. Ajmera. Convolutional neural network-based human identification using outer ear images. In *Soft computing for problem solving*, pages 707–719. Springer, 2019.

[29] D. Štepec, Ž. Emeršič, P. Peer, and V. Štruc. Constellation-based deep ear recognition. In *Deep biometrics*, pages 161–190. Springer, 2020.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] M. Vitek, P. Rot, V. Štruc, and P. Peer. A comprehensive investigation into sclera biometrics: a novel dataset and performance study. *Neural Computing and Applications*, 32(24):17941–17955, 2020.

[32] X. Wu, E. Boutellaa, X. Feng, and A. Hadid. Kinship verification from faces: Methods, databases and challenges. In *ICSPCC*, pages 1–6, 2016.

[33] X. Wu, E. Granger, T. H. Kinnunen, X. Feng, and A. Hadid. Audio-visual kinship verification in the wild. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[34] Y. Wu, Z. Ding, H. Liu, J. Robinson, and Y. Fu. Kinship classification through latent adaptive subspace. In *IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 143–149, 2018.

[35] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[36] J. Yu, M. Li, X. Hao, and G. Xie. Deep fusion siamese network for automatic kinship verification. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020.

[37] Žiga Emeršič, A. S. V. Kumar, B. S. Harish, W. Gutfeter, J. N. Khiarak, A. Pacut, E. Hansley, M. P. Segundo, S. Sarkar, H. Park, G. P. Nam, I. J. Kim, S. Sangodkar, U. Kacar, M. Kirci, L. Yuan, J. Yuan, H. Zhao, F. Lu, J. Mao, X. Zhang, D. Yaman, F. I. Eyiokur, K. B. Ozler, H. K. Ekenel, D. P. Chowdhury, S. Bakshi, P. K. Sa, B. Majhni, P. Peer, and V. Štruc. The unconstrained ear recognition challenge 2019. In *International Conference on Biometrics (ICB)*, 2019.