



## UvA-DARE (Digital Academic Repository)

### Vision based kinship recognition

Wang, W.

**Publication date**

2023

**Document Version**

Final published version

[Link to publication](#)

**Citation for published version (APA):**

Wang, W. (2023). *Vision based kinship recognition*. [Thesis, fully internal, Universiteit van Amsterdam].

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.



# Vision Based Kinship Recognition



wang wei

Vision Based Kinship Recognition

Wei Wang





# Vision Based Kinship Recognition

This book was typeset by the author using L<sup>A</sup>T<sub>E</sub>X 2<sub>ε</sub>.

Cover drawing: Chenlei Li  
Cover design: Chenlei Li & Wei Wang

Copyright © 2023 by Wei Wang.

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the author.

ISBN 978-90-9037301-0



# Vision Based Kinship Recognition

## ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor  
aan de Universiteit van Amsterdam  
op gezag van de Rector Magnificus  
prof. dr. ir. P.P.C.C. Verbeek  
ten overstaan van een door het college voor promoties ingestelde commissie,  
in het openbaar te verdedigen in de Agnietenkapel  
op dinsdag 16 mei 2023, te 12.00 uur

door

**Wei Wang**

geboren te Shandong

*Promotiecommissie*

Promotor:	prof. dr. T. Gevers	Universiteit van Amsterdam
Copromotores:	dr. S. You	Universiteit van Amsterdam
	dr. S. Karaoglu	Universiteit van Amsterdam
Overige leden:	prof. dr. M. Worring	Universiteit van Amsterdam
	prof. dr. A. A. Salah	Universiteit Utrecht
	prof. dr. ir. P. H. N. de With	Technische Universiteit Eindhoven
	dr. A. Visser	Universiteit van Amsterdam
	dr. E. Gavves	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica



UNIVERSITEIT VAN AMSTERDAM

The research was supported by China Scholarship Council (CSC) under project number 201807930015. The work described in this thesis has been carried out at the Computer Vision Lab of the University of Amsterdam.



---

## CONTENTS

---

1	INTRODUCTION	1
1.1	Research Outline and Questions . . . . .	3
1.2	Societal Impact and Ethical Statement . . . . .	6
1.3	Origins . . . . .	7
2	A SURVEY ON KINSHIP VERIFICATION	9
2.1	Introduction . . . . .	9
2.2	Motivation and Background . . . . .	10
2.2.1	Biological Background . . . . .	10
2.2.2	Kinship Terminology, Types and Classes . . . . .	11
2.2.3	Kinship Information . . . . .	13
2.2.4	Challenges on Kinship Datasets . . . . .	13
2.2.5	Architecture of Kinship Verification Systems . . . . .	14
2.3	Datasets . . . . .	14
2.3.1	Kinship Datasets: 4-types . . . . .	15
2.3.2	Kinship Datasets: 7-types . . . . .	17
2.3.3	Kinship Datasets: 11-types . . . . .	18
2.3.4	Others . . . . .	18
2.3.5	Discussion . . . . .	19
2.4	Representative Methods . . . . .	19
2.4.1	Image-based . . . . .	21
2.4.2	Video-based . . . . .	30
2.4.3	Multi-label Methods . . . . .	31
2.4.4	Discussion . . . . .	32
2.4.5	Potential Directions for Kinship Verification . . . . .	33
2.5	Nemo-dataset . . . . .	34
2.5.1	Motivation . . . . .	34
2.5.2	Data Collection . . . . .	34
2.5.3	Data Statistics . . . . .	35
2.6	Evaluation Protocols, Metrics for Kinship Verification . . . . .	36
2.6.1	Protocols . . . . .	36
2.6.2	Metrics . . . . .	36
2.7	Benchmarking . . . . .	37
2.7.1	Benchmark on Publicly Available Datasets . . . . .	39
2.7.2	Benchmark on the Nemo-Kinship Dataset . . . . .	41
2.7.3	Discussion . . . . .	44
2.8	Conclusion . . . . .	44
3	KINSHIP IDENTIFICATION THROUGH JOINT LEARNING	45
3.1	Introduction . . . . .	45



## CONTENTS

3.2	Related Work . . . . .	46
3.3	Kinship Identification through Joint Learning with Kinship Verification . . . . .	47
3.3.1	Definition of Kinship Verification, Kinship Identification and Kinship Classification . . . . .	47
3.3.2	Relationship between Kinship Verification and Kinship Identification and the Limitation of Existing Methods . . . . .	48
3.4	Joint Learning of Kinship Identification and Kinship Verification . . . . .	50
3.4.1	Architecture of the Proposed Joint Learning Network (JLNet) . . . . .	50
3.4.2	Comparative Methods . . . . .	52
3.5	Experiments . . . . .	53
3.5.1	Unbias Dataset for Training and Testing . . . . .	53
3.5.2	Experimental Design . . . . .	53
3.5.3	Results & Evaluation . . . . .	54
3.6	Conclusion . . . . .	57
4	IDENTITY INVARIANT AGE TRANSFER FOR KINSHIP VERIFICATION OF CHILD-ADULT IMAGES . . . . .	59
4.1	Introduction . . . . .	59
4.2	Related Work . . . . .	60
4.2.1	Kinship Verification . . . . .	60
4.2.2	Age-Invariant Face Feature Learning and Cross-Age Face Synthesis . . . . .	61
4.3	Method . . . . .	62
4.3.1	Problem Formulation and Motivation . . . . .	62
4.3.2	Pipeline . . . . .	62
4.3.3	Identity-preserved Aging Generator . . . . .	62
4.3.4	Identity-Invariance-Aging-Transferring Network . . . . .	64
4.3.5	Nemo-Kinship-Children Dataset . . . . .	65
4.4	Experiments . . . . .	66
4.4.1	Data Selection and Preparation . . . . .	66
4.4.2	Experimental Setups . . . . .	66
4.4.3	Comparison with Current Methods . . . . .	67
4.4.4	Ablation Study . . . . .	69
4.4.5	Robustness and Generalization . . . . .	70
4.5	Conclusion . . . . .	70
5	KINSHIP SIMILARITY FOR OPEN SETS . . . . .	71
5.1	Introduction . . . . .	71
5.2	Related Work . . . . .	72
5.2.1	Kinship Recognition and Related Tasks . . . . .	72
5.2.2	Open-set Recognition . . . . .	73
5.3	Problem Formulation and Comparison . . . . .	73
5.3.1	Problem Formulation . . . . .	73
5.3.2	Comparison with Kinship Recognition . . . . .	74
5.3.3	Comparison with Open-set Recognition . . . . .	75
5.4	Methodology . . . . .	76
5.4.1	Hierarchical Information . . . . .	76

5.4.2	Distance-based Losses . . . . .	77
5.5	Experiment . . . . .	78
5.5.1	Datasets . . . . .	78
5.5.2	Experimental Settings . . . . .	79
5.6	Conclusion . . . . .	83
6	KINSHIP VERIFICATION IN VIDEOS USING SEMI-SUPERVISED LEARNING	85
6.1	Introduction . . . . .	85
6.2	Related Work . . . . .	86
6.2.1	Kinship Verification . . . . .	86
6.2.2	Transformer . . . . .	87
6.2.3	Pre-training without Annotation . . . . .	87
6.3	Method . . . . .	88
6.3.1	Problem Formulation . . . . .	88
6.3.2	Feature Learning through Video-kin Augmentation . . . . .	88
6.3.3	Kinship-transformer . . . . .	89
6.3.4	Pre-training of Kinship-transformer . . . . .	90
6.3.5	Fine-tuning of Kinship-transformer on Kinship Related Datasets	91
6.3.6	Similarity during Testing . . . . .	91
6.4	Experiments . . . . .	91
6.4.1	Datasets . . . . .	91
6.4.2	Implementation Details . . . . .	92
6.4.3	Results . . . . .	93
6.4.4	Ablation Study . . . . .	95
6.5	Conclusion . . . . .	98
7	SUMMARY AND CONCLUSION	99
7.1	Summary . . . . .	99
7.2	Conclusion . . . . .	101
A	APPENDIX	103
A.1	Software & Repositories . . . . .	103
	Bibliography	118
	Samenvatting	119
7.2	Samenvatting . . . . .	119
7.3	Gevolgtrekking . . . . .	121
	Acknowledgments	123





---

## INTRODUCTION

---

Homines non nascentur, sed finguntur

*Desiderius Erasmus*

In general, people tend to live in groups. This group bond between people has fostered the development of language, culture, science, and society structures. Since humans belong to the group of mammals, raising babies and children is enshrined in their care, interaction, and communication skills. Humans have kinship and family ties from the existence as hunters/gatherers, the transition to agricultural workers up to an industrial and modern society.

Today, we live in an interactive and communicative world like never before. Despite the abundance and ever-changing social communication and connections, one social trait has remained the same: kinship. In fact, kinship and kinship recognition are essential components of human social relations. Family affection influences our growth, environment, character, development, and social status.

In many cultures, artistic expressions are closely tied to kinship relationships. Figure 1 (a family oil portrait by Rembrandt) shows the strong relationships between family and art. Kinship is also related to religious beliefs and practices [94, 186], as many religions include specific rules and customs for marriage, family, and inheritance [58]. Moreover, the intrinsic ability of humans for kinship recognition has an impact on both direct fitness (breeding behavior) and indirect fitness (altruistic behavior) [93]. In addition, kinship and the recognizing of kinship are of great importance to today's society [64, 93, 146], including finding missing children, genealogy research [43, 95], forensic research, social behavior analysis [64, 93, 146, 240], and so forth. With the rise of computers, computer vision and learning systems, the question is whether technology can help us in analyzing kinship and recognizing it. This is also known as vision-based kinship recognition [6, 165].

Vision based kinship recognition is a technology that automatically recognizes familial relationships based on visual information [6, 165]. Previous research [9, 38, 93, 154] has shown that there exists a relationship between face similarity and kinship. This gives us the opportunity to conduct vision-based kinship recognition using facial information. Figure 2 shows the representative subtasks in the kinship recognition field, using facial information.



Figure 1: "Brunswick Family Portrait" by Rembrandt Harmenszoon van Rijn.

In the past decade, technological advancements have been made in the vision based kinship recognition field [156, 165, 166, 170, 214, 217]. Facial based kinship datasets are collected [61, 62, 129, 158, 163, 216, 218], and different kinship recognition methods are proposed. In the early days of kinship recognition, local features corresponding to eyes and noses are often used, as they show heritable similarities. Hand-crafted feature extractors like HOG [207], LBP [3], and Gabor [1, 133] are commonly employed [130, 158, 159, 176, 227, 228, 251]. As the field has evolved, metric learning has become a widely used technique for kinship representation. Metric learning aims to improve kinship representation by maximizing the inter-class and minimizing the intra-class distributions [70, 157, 227, 254]. Methods like NRML [131], LMMML [86], and DMML [229] are proposed. Recently, the focus is on deep learning-based methods [7, 30, 37,

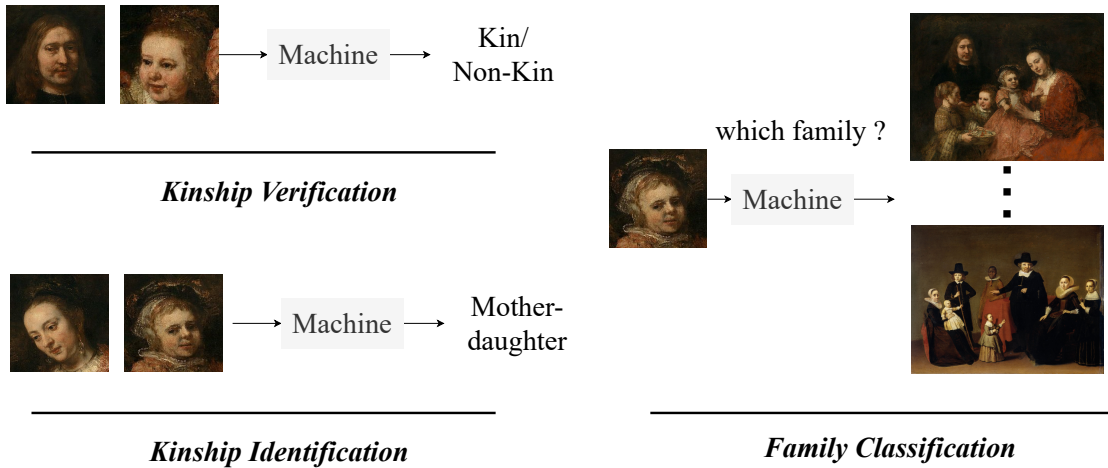


Figure 2: Representative subtasks in kinship recognition.

54, 119, 142, 180, 195, 199, 232, 238]. For instance, graph-based kinship reasoning (GKR) [114] fuses the extracted features using Graph Neural Networks. Supervised Mixed Norm AutoEncoder (SMNAE) [100] learns spatio-temporal representation by using autoencoders. As such, deep learning-based kinship recognition is the preferred choice for kinship recognition research and development.

Despite the progress made in the field of kinship recognition, there are still a few shortcomings. This is due to a combination of intrinsic challenges related to the nature of the task itself and extrinsic (imaging) challenges:

- Intrinsic challenges include the age, gender, expression, ethnicity, and types of genetic relationships of the individuals being recognized. These factors lead to variations in the appearance of facial features.
- Extrinsic challenges include variations in imaging conditions, such as lighting and resolution, as well as biased testing datasets that do not accurately represent the diversity of the population. There may also be unknown classes of kinship relationships that are not accounted for in the current recognition methods, limiting the generalizability of these techniques.

These challenges are common in real scenarios. However, not all of these challenges are well-studied. Consequently, given these difficulties, this thesis aims to investigate how to perform kinship identification kinship recognition in more realistic, real-world settings.

## 1.1 RESEARCH OUTLINE AND QUESTIONS

To study kinship recognition, we focus on the following question:

### ***What is kinship verification and what are the challenges?***

As a starting point, we focus on kinship verification since it is a well-studied task in kinship recognition. Kinship verification is defined as the automatic process of verifying whether two or more persons are blood relatives (kin) by analyzing images of their faces.

Kinship verification is an important research field in computer vision with many applications such as finding missing persons, family album organization, and online image search. Although substantial progress has been made in kinship verification, there are still challenges such as intrinsic (face *i.e.*, differences in facial appearance) and extrinsic (acquisition *i.e.*, varying imaging conditions) problems. Moreover, there is still a demand for more diverse kinship datasets. Reviewing the existing literature in kinship verification provides a holistic understanding of challenges in kinship recognition.

Therefore, Chapter 2 provides a survey on kinship verification methods and datasets. The survey starts with the definition of kinship verification and its corresponding intrinsic and extrinsic challenges. Then, an overview of kinship verification methods and datasets is given. Finally, a new multi-modal dataset (Nemo-Kinship Dataset) is proposed as a benchmark dataset addressing large inter-subject age variations consisting of 4216 videos



of 248 persons from 85 families. The newly collected dataset is used to systematically test and analyze state-of-the-art methods.

In Chapter 2, we explore the relationship between the tasks of kinship verification, identification, and family classification. Kinship verification is a well-explored task: identifying whether two persons are kin. In contrast, kinship identification has been largely ignored so far. Kinship identification aims to further identify the particular type of kinship. An extension to kinship verification falls short to properly obtain identification because existing verification networks are individually trained on specific kin types and do not consider the context between different kinship types. Also, existing kinship verification datasets have biased positive-negative distributions, which are different from real-world distributions. Many experiments done so far use equal numbers of positive and negative examples. However, in real scenarios, the number of negative samples is usually larger than the number of positive ones. Therefore, the second research question is:

***How can we better verify kin-types when facing unbalanced distributions in real scenarios?***

To address this question, in Chapter 3, we propose a novel kinship identification approach based on joint training of kinship verification ensembles and classification modules. We propose to rebalance the training dataset. Large scale experiments demonstrate the improved overall performance on kinship identification. The experiments further show a performance improvement of kinship verification when trained on the same dataset with more realistic distributions.

In addition to the extrinsic challenges of unbalanced data, intrinsic challenges can also negatively influence the performance of kinship recognition algorithms. Aging is one of the intrinsic challenges which may cause a change in facial appearance. Therefore, the third research question is:

***How can we alleviate the negative influence of age variations when conducting kinship verification?***

The performance of kinship verification can be influenced by children-adult pairs due to the large variations in facial appearance and shape. In Chapter 4, this problem is approached from an age-transferring generative modeling perspective, and we present a unified approach to kinship verification for child-adult pairs. Kinship features are computed with the aim to eliminate the discrepancy between the features of children and adults through age transferring.

To this end, a Children-Adult-Transferring (CAT) Module is proposed by exploiting the generative knowledge obtained by an Identity-Preserved Conditional Generator. In fact, children’s images are generated from the childhood age domain to an adulthood domain, and the latent feature through generation is utilized. Then, a Kinship Mapping Module (KMM) is created for mapping the latent features to the kinship-related domain, which is further handled by the Neighborhood Repulsed Metric Learning method. Since

there is no public kinship verification dataset containing a large variety of children-adult images, the Nemo-Kinship-Children dataset is created. The experimental results show that the towards-adult transferred features of children images robustly represent kinship relations.

While kinship recognition and related methods have a wide range of applications in computer vision, many of these approaches are focused on closed sets, where the number of possible relationships is limited. However, in real-life scenarios, kinship recognition is an open set problem as there are many unknown and unlabeled kin classes. This brings us to the following question:

***How can we improve the kinship recognition when facing unknown classes?***

To address the challenge of kinship recognition in open set scenarios, Chapter 5 focuses on the measurement of kinship at various degrees for collections that include both kin and non-kin-related people. The aim is to determine family relationships and their corresponding degrees of kinship hierarchically. To achieve this, we propose a novel and more general task called the Open-set Kinship Similarity Measurement (OKSM).

Different from ordinary open set methods, our method is pairwise-based and is able to exploit mutual information from positive pairs. Large scale experiments and ablation studies show that our method (1) reaches SOTA performance on the FIW dataset in open set, (2) is able to properly separate kinship categories using pairwise similarity, and (3) generates uniform similarity distributions.

Another major challenge in the field of kinship recognition is the limited size of available datasets. It is often due to the tedious (labor-consuming) process of annotating kinship relationships. To deal with small datasets, transfer learning can be used for kinship verification by exploiting off-the-shelf facial knowledge. However, how to learn kinship distributions without annotations has not been considered before. As a result, the fifth research question is:

***How can we explore off-the-shelf knowledge from pretrained facial networks for kinship verification with limited kinship datasets?***

In Chapter 6, a two-branch model, Kinship-transformer (KT), is proposed in a semi-supervised manner. The model is first pre-trained on a large collection of images containing faces (without kinship labels). Then the model is fine-tuned on a small video kinship dataset (Nemo-kinship). To achieve this, we propose a kinship-oriented augmentation method (Video-kin augmentation) for pretraining. During the pretraining, the original video is augmented into different styles, trying to form feature representations that are similar to the kinship distribution. The results show the superiority of the proposed model compared to traditional convolutional neural networks.

## 1.2 SOCIETAL IMPACT AND ETHICAL STATEMENT

Vision-based kinship recognition is closely related to our lives and has many potentially valuable applications. The technology can potentially benefit our society by enabling faster and more accurate identification of family members in emergencies and reuniting separated families in the event of a kidnapping or disaster [36]. However, at the same time, it also raises ethical and social concerns.

One of the major ethical concerns of vision-based kinship recognition is the potential for invasion of privacy. Because kinship recognition uses an individual's facial features and automatically identifies/verifies family relationships based on facial information, sensitive information can be used without people's consent. Such sensitive information can potentially be used for mass surveillance and tracking of individuals and families. This can raise concerns about individual privacy violations.

Another ethical concern associated with vision-based kinship recognition is the potential for discrimination and bias. Kinship recognition algorithms may be biased, resulting in inaccurate or unfair results. For example, facial recognition algorithms may be more likely to misidentify people of certain races, genders, or ages [67], which may lead to discriminatory results in kinship recognition.

Overall, the social implications and ethical issues concerning vision-based kinship recognition are multifaceted. While this technology has the potential to provide many benefits, it also raises important concerns about privacy, discrimination, and the potential for increased surveillance and control. It is crucial for the researchers developing and implementing kinship recognition to consider and address these ethical issues carefully. To ensure its responsible and fair use, we obey the following ethical principles:

- The research participants are fully informed about the purpose of the research and how their biometric data is being collected and used [248].
- Individuals have the right to control their own biometric data and the right to consent or disagree with its use [24, 173].
- The use of kinship recognition technologies should aim to maximize benefits and minimize harm to individuals and society [11, 138].
- The personal and family information of the participants in kinship recognition experiments are kept and protected confidentially [71, 185].
- The kinship recognition technology should not be used in a way that causes harm or distress to individuals [248].
- The use of kinship recognition technology should remain as fair and impartial as possible, avoiding bias and discrimination against marginalized groups [160].

In conclusion, the use of kinship recognition techniques raises important ethical considerations. By addressing these challenges and considering the potential impact on society, the researchers involved in developing and implementing this technology can help ensure that it is used ethically and effectively.



## 1.3 ORIGINS

This thesis is based on the following works:

- **Chapter 2** is based on "A Survey on Kinship Verification", published in *Neuro-computing*, 2023, by Wei Wang, Shaodi You, Sezer Karaoglu, Theo Gevers.

*Contribution of authors*

Wei Wang:	Methodology, data collecting, experiments, writing,
Shaodi You:	Resources, methodology, data collecting, supervision, writing - review & editing,
Sezer Karaoglu:	Resources, methodology, data collecting, supervision, writing - review & editing,
Theo Gevers:	Resources, conceptualization, methodology, formal analysis, investigation, writing – review & editing, supervision.

- **Chapter 3** is based on "Kinship Identification through Joint Learning using Kinship Verification Ensembles", published in *European Conference on Computer Vision*, 2020, by Wei Wang, Shaodi You, Sezer Karaoglu, Theo Gevers.

*Contribution of authors*

Wei Wang:	Methodology, experiments, writing,
Shaodi You:	Resources, conceptualization, methodology, formal analysis, supervision, writing - review & editing,
Sezer Karaoglu:	Resources, conceptualization, supervision, writing - review & editing,
Theo Gevers:	Resources, conceptualization, methodology, formal analysis, investigation, writing – review & editing, supervision.

- **Chapter 4** is based on "Identity Invariant Age Transfer for Kinship Verification of Child-Adult Images", under review in *Computer Vision and Image Understanding*, 2022, by Wei Wang, Shaodi You, Yahui Zhang, Sezer Karaoglu, Theo Gevers.

*Contribution of authors*

Wei Wang:	Conceptualization, methodology, experiments, writing,
Shaodi You:	Resources, conceptualization, methodology, formal analysis, supervision, writing - review & editing,
Yahui Zhang:	Data collecting,
Sezer Karaoglu:	Resources, conceptualization, methodology, data collecting, supervision, writing - review & editing,
Theo Gevers:	Resources, conceptualization, methodology, formal analysis, investigation, writing – review & editing, supervision.

- **Chapter 5** is based on "Kinship Similarity for Open Sets", under revision in *Pattern Recognition*, 2023, by Wei Wang, Shaodi You, Sezer Karaoglu, Theo Gevers.

*Contribution of authors*

Wei Wang:	Conceptualization, methodology, experiments, writing,
Shaodi You:	Resources, methodology, conceptualization, formal analysis, supervision, writing - review & editing,
Sezer Karaoglu:	Resources, methodology, supervision, conceptualization, supervision, writing - review & editing,
Theo Gevers:	Resources, conceptualization, methodology, formal analysis, investigation, writing – review & editing, supervision.

- **Chapter 6** is based on "Kinship Verification in Videos using Semi-Supervised Learning", under review in *International Conference on Computer Vision*, 2023, by Wei Wang, Shaodi You, Sezer Karaoglu, Theo Gevers.

*Contribution of authors*

Wei Wang:	Methodology, experiments, writing,
Shaodi You:	Resources, methodology, formal analysis, supervision, writing - review & editing,
Sezer Karaoglu:	Resources, methodology, supervision, supervision, writing - review & editing,
Theo Gevers:	Resources, conceptualization, methodology, formal analysis, investigation, writing – review & editing, supervision.

The author has further contributed to the following publications:

- Le Minh Ngo, Wei Wang, Burak Mandira, Sezer Karaoglu, Henri Bouma, Hamdi Dibeklioglu, Theo Gevers, Identity Unbiased Deception Detection by 2D-to-3D Face Reconstruction, published in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.
- Jian Han, Wei Wang, Sezer Karaoglu, Wei Zeng, Theo Gevers, Pose invariant age estimation of face images in the wild, published in *Computer Vision and Image Understanding*, 2021.

---

## A SURVEY ON KINSHIP VERIFICATION

---

### 2.1 INTRODUCTION

Kinship verification (KV) is defined as the automatic process of verifying whether two or more persons, represented by images of their faces, are blood relatives *i.e.*, kin or no-kin [156, 166, 170, 214, 217]. Image-based kinship verification assumes facial resemblances between genetic-related persons [40, 93, 240]. Fang *et al.* [62] are among the first to study kinship verification based on images of faces. Since then, kinship verification attracted a lot of attention in computer vision and related research fields such as historical and genealogical studies [43, 95], social media [20, 43, 236], behavior analysis [64, 93, 146, 240], and inheritance [165]. Kinship verification is a challenging task mainly due to intrinsic (face *i.e.*, differences in facial appearance) and extrinsic (acquisition *i.e.*, varying imaging conditions) challenges. Intrinsic challenges are related to changes in age, gender, expression, ethnicity, and types of genetic relationships [156]. Extrinsic challenges correspond to the image acquisition process, such as changes in illumination, camera viewpoint, and face occlusion.

Kinship verification can be divided into three groups based on the process of feature extraction and learning: (1) (hand-crafted) feature extraction, (2) metric learning, and (3) deep learning. Early kinship verification methods focus on extracting features at facial landmarks such as eyes and noses. Hand-crafted descriptors include HOG [39], LBP [147], PEM [109], and Gabor [1] features. Later, metric learning methods are proposed to exploit (distance) metrics by maximizing inter-class and minimizing intra-class distances. More recently, deep learning is proposed to learn features and metrics simultaneously [136, 192].

Different image datasets are proposed. The CornellKin dataset, proposed by Fang *et al.* [62] in 2010, is the first widely used image dataset. Then, the KinFaceW-I & II [129, 131] public datasets are proposed containing four different kinship types (father-son, mother-son, father-daughter, and mother-daughter). Robinson *et al.* propose the Families In the Wild (FIW) dataset [166, 168] to study kinship verification in more challenging and dynamic environments. In addition, a number of video datasets are provided [49, 100, 226]. Unfortunately, the major problem with these datasets remains the limited age range between subjects.

To this end, in this chapter, we propose a multi-modal dataset for kinship verification containing a wider range of age variations than existing datasets. The newly collected Nemo-kinship dataset consists of 4216 videos of 85 families with 248 individuals.

This survey:

- provides a large survey on kinship verification methods and datasets.
- studies the challenges of existing kinship methods and discusses future directions.
- proposes the Nemo-Kinship dataset containing a large range of age differences between subjects.

This survey is organized as follows. In Section 2, kinship verification is discussed including kinship definition, biological background, and potential applications. An overview is given of different kinship datasets and methods in Sections 3 and 4, respectively. In Section 5, the Nemo-kinship dataset is presented. Evaluation protocols for kinship verification are given in Section 6. In Section 7, a benchmark is conducted on both public datasets as well as on the Nemo-kinship dataset. Conclusion, discussion, and future directions are outlined in Section 8.

## 2.2 MOTIVATION AND BACKGROUND

### 2.2.1 *Biological Background*

#### *Kinship verification by humans*

Facial information is the most commonly used identification cue in genetic similarity [9,38,93,154,240]. Images of faces contain important identification cues to determine, for example, the age, identity, gender, and ethnicity of a person [93,236]. In 1982, Daly and Wilson [40] propose to use facial similarity as a physiological cue for kinship detection, providing a basis for human kinship detection [99]. Moreover, kinship verification is used to measure direct (breeding behavior) and indirect fitness (altruistic behavior) [93]. For instance, the paternal resemblance [20] has a positive effect on family relationships, and the facial resemblance enhances the corporation as well as trust [42,154].

#### *Significance and applications*

The above factors indicate that kinship verification is beneficial for genealogical studies, but it also has important implications on other applications such as arranging and managing hundreds of thousands of images online [182], historic lineage and genealogical studies identifying inaccessible people based on their kinship similarity. Moreover, in forensic and criminal studies, kinship verification is used to reduce the number of suspects by narrowing down the search space *e.g.*, in the case of the Boston Marathon bombing [97]. Hence, kinship verification may have a positive influence on different domains such as genealogical studies, social media, and forensic investigation, with many applications such as automatic photo tagging and management, missing children, crime scene investigation, and surveillance. However, improper use of kinship verification can lead to privacy violations. Moreover, the verification system's security may fail in case of adversarial attacks [72,75,177,178,247] and fake facial images [96,102].

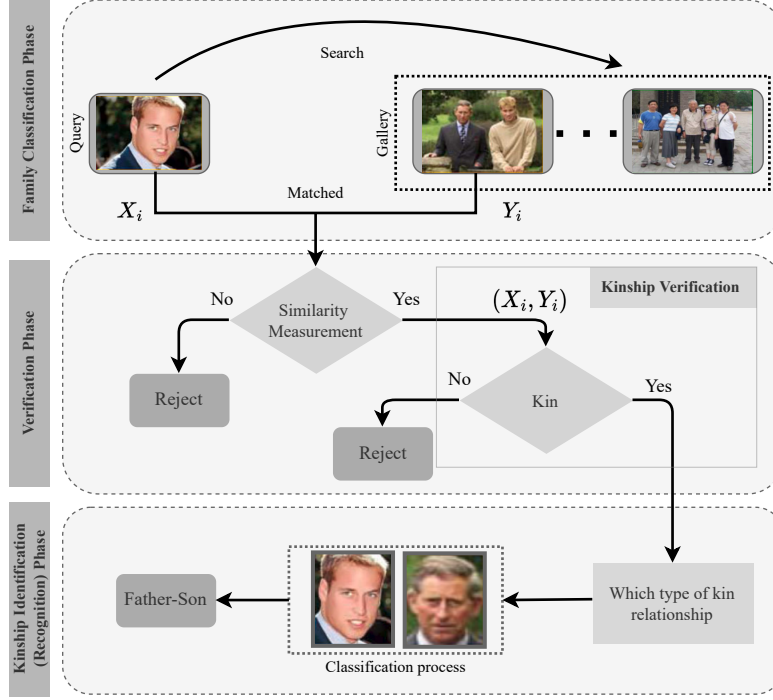


Figure 3: Flowchart of the three different tasks in the domain of kinship recognition.

### 2.2.2 Kinship Terminology, Types and Classes

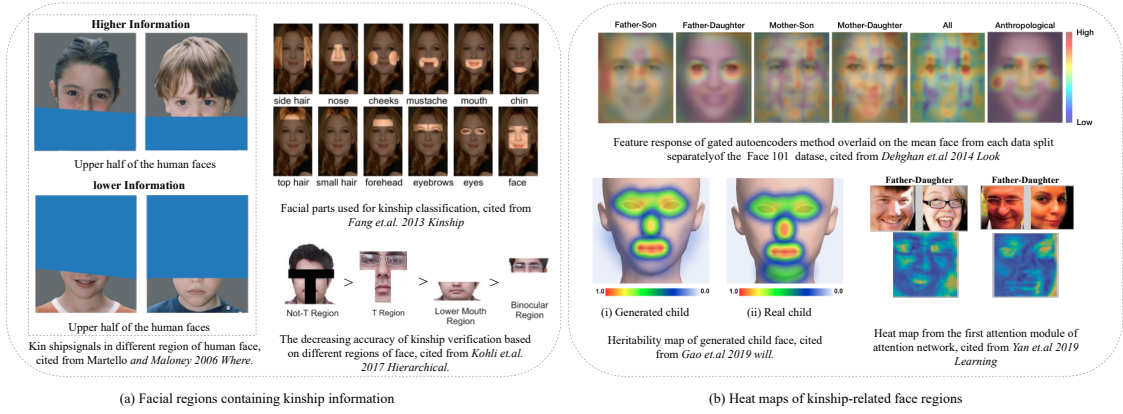
#### Kinship verification, recognition and identification

Mohammed *et al.* [6],

*In general, kinship may indicate similarity, familiarity, or closeness between entities on the basis of some or all of the basic traits or features ... biology, kinship typically refers to the degree of genetic relatedness or coefficient of relationship between individual members of the same species [16, 39, 118]*  
...

In general, there are two kinship types: kinship with blood (consanguineal kinship) and marriage ties (affinal kinship) [210]. Kinship with blood ties corresponds to blood-based relationships with overlapping genes [41] and kinship with marriage ties addresses the connection based on marriage. This chapter focuses on blood ties. According to the degree of similarity between family members, kinship is classified into three groups [6, 140, 141]:

1. Primary kinship: Blood ties between people do not contain intermediate relationships. This class consists of Father-Daughter (F-D), Father-Son (F-S), Mother-Daughter (M-D), Mother-Son (M-S), Sister-Brother (S-B), Sister-Sister (S-S), and Brother-Brother (B-B) (ego-self and affinal relationships are not considered).
2. Secondary kinship: Blood ties between people contain one intermediate kinship person, *e.g.*, Uncle-Nephew. Common relationships are GrandFather-GrandDaughter (GF-GD), GrandFather-GrandSon (GF-GS), GrandMother-GrandDaughter (GM-GD), and GrandMother-GrandSon (GM-GS).



(a) *Facial regions containing kinship information.*

(b) Heat maps of kinship regions.

Figure 4: Facial regions used for kinship verification and kinship-related heat maps.

3. Tertiary kinship: Two intermediate levels exist for this class of kinship [141].

Hence, the type of kinship relationship can be categorized into 4-types (parents-children included), 7-types (siblings included), and 11-types (grandparents-grandchildren included).

### *Kinship Verification: definition*

According to [6, 166], kinship recognition is the task of studying blood relationships based on facial image information. Kinship verification is one of the subtasks of kinship recognition. The three major subtasks of kinship recognition are [6, 166]: (1) kinship verification defined as a binary classification task determining whether two or more persons are blood-related, (2) kinship identification with the aim to estimate the kin-type, and (3) kinship/family classification identifying to which family an individual belongs to [76, 166, 217]. These three tasks are interrelated and influence each other [6]. As shown in Figure 3, kinship verification is based on the results generated by kinship classification. Furthermore, kinship verification analyzes different types of kinship relationships [6]. Hence, kinship verification plays a central role in kinship recognition.

## Formulation

As discussed in *Section. 2.2.2*, kinship verification is a binary classification task determining whether two (or more) people are kin or not. We now briefly discuss the formalized kinship verification task [114]. Most of the existing research focuses on bi-subject (one-versus-one) kinship verification. A canonical definition of the task is as follows: let  $\mathcal{P} = \{(X_i, Y_i) \mid i = 1, 2, \dots, N\}$  denote the training set of images pairs containing kin relationships for each kin-type.  $N$  is the number of positive pairs, and  $X_i$  and  $Y_i$  are parent and children images respectively. Then, let the negative training set be denoted by  $\mathcal{N} = \{(X_i, Y_j) \mid i, j = 1, 2, \dots, N, i \neq j\}$ , representing the image pairs without kin relation. To verify the kin-types, the binary classifier  $f(\cdot)$  and feature extractor  $g(\cdot)$  are used. Then, the final output is formulated by:

$$\mathbf{z} = f(g(\mathbf{X}_i, \mathbf{Y}_j)), \mathbf{z} \in \{0, 1\}, \quad (1)$$



where 1 represents kin and 0 non-kin. There are special cases where two parents and a child are used as input. For this tri-subject (one-versus-two) kinship verification task, the positive training set is given by  $\{(X_{fi}, X_{mi}, Y_{ci}) \mid i = 1, 2, \dots, N\}$  and the negative training set is denoted by  $\{(X_{fi}, X_{mi}, Y_{cj}) \mid i, j = 1, 2, \dots, N, i \neq j\}$ . Then, the final output is given by:

$$z = f(g(X_{fi}, X_{mj}, Y_{cj})), z \in \{0, 1\}, \quad (2)$$

where  $X_{fi}, X_{mi}, Y_{ci}$  denote the  $i$ th sample of father, mother and child.

### 2.2.3 Kinship Information

Obviously,  $f(\cdot)$  needs to make full use of kinship information from  $X_i$  and  $Y_j$ . However, how to effectively extract kinship information is still a question. Martello and Maloney [38] conduct experiments with 220 participants. They show that the upper half of the face contains more relevant kinship information than the lower half. Furthermore, eye regions contain slightly more useful cues than the rest of the upper half of the face. Hence, enhancing the eyes, nose, and mouth areas may improve the accuracy of kinship verification. Studies [62, 66, 99, 232] also show that cues, related to kinship information, can be based on machine-based kinship verification. However, there are conflicting findings between different studies. Gao *et al.* [66] show that mouth regions contain higher similarities between children and parents. In contrast, Martello and Maloney [38] show that people are better at predicting kinship without mouth regions. In addition, DeBruine *et al.* [43] show that the degree of similarity may vary between same-gender and different-gender pairs. The same-gender pairs usually obtain higher similarities. Figure 4b shows the important facial cues for different kin-types [44]. Features may vary for different kin-types [165].

### 2.2.4 Challenges on Kinship Datasets

As mentioned in Section 2.1, there are different intrinsic and extrinsic challenges. Compared to datasets for face recognition [136, 193], kinship datasets are much smaller in size. Hence, new kinship datasets are required according to:

- Large-scale video-based kinship datasets.
- Kinship datasets for solving specific kinship-related problems.

In general, current kinship datasets consist of still images. However, video-based datasets contain dynamic facial and head cues, including head motion (gait), expressions and mouth movement. Video-based datasets may increase the accuracy and robustness of kinship verification algorithms. Another important aspect of a kinship dataset is that it can be used for solving specific kinship relationships. For example, the face of a person changes over time (*i.e.*, aging) and may negatively influence existing kinship verification methods. Therefore, a dataset containing pictures/videos of the same person over time is an important addition.

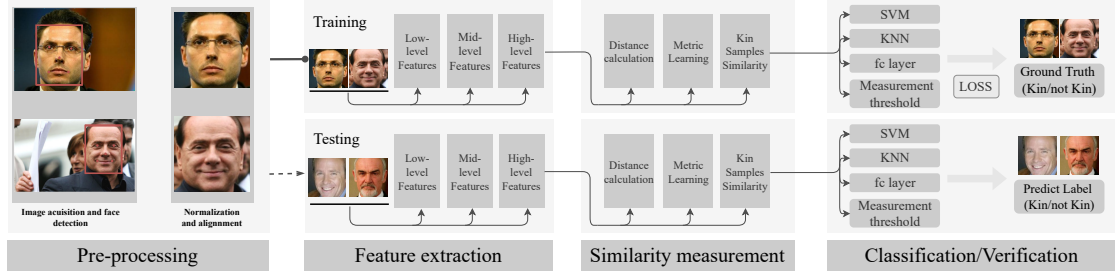


Figure 5: Pipeline of the kinship verification system.

### 2.2.5 Architecture of Kinship Verification Systems

Studies [6,41,125,156] show that an automated kinship verification system can be divided into four phases: (1) face detection, (2) feature extraction, (3) similarity computation, and (4) verification. The pipeline of a kinship verification system is illustrated in Figure 5.

#### *Pre-processing phase*

The pre-processing phase locates, detects, and segments the facial regions and separates them from the background. It ensures that the kinship verification system focuses on valuable regions to extract features. This phase also includes the normalization of head pose, illumination, and scale.

#### *Feature extraction phase*

Feature extraction methods are proposed based on hand-crafted descriptors such as texture, appearance, and geometry features. Other feature extraction methods employ deep neural networks.

#### *Similarity measurement phase*

This phase measures the similarity between image pairs based on the extracted features. It includes selecting the best subset from the obtained feature or mapping the extracted features to a more prominent manifold. Different distance calculations (*e.g.*, Euclidean and cosine distance) together with metric learning are used.

#### *Verification phase*

The verification phase outputs the final result *i.e.*, kin or non-kin. Commonly conventional machine learning methods that are used are SVM and KNN. For deep learning methods, the classification results are usually obtained through the *fc* layer or MLP.

## 2.3 DATASETS

Fang *et al.* [62] collect the first kinship dataset. Since then, different datasets are collected to narrow the distribution discrepancy between training and real-world data. Increasingly

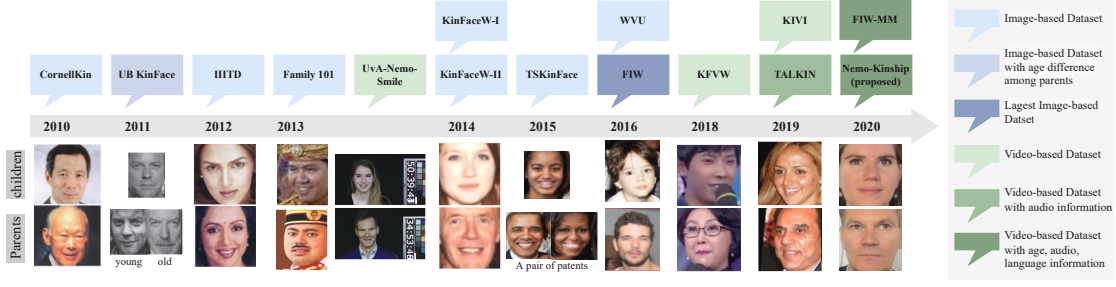


Figure 6: Development of representative kinship datasets for kinship verification.

Table 1: Checkerboard of existing datasets. The darker the box is, the more similar.

kin types	constrained	unconstrained	pairs <300	pairs 300~500	pairs 500~1000	pairs >1000	videos	audio	family structure
2		family 101 [61] TSKinFace [158]			family 101 [61]	TSKinFace [158]			family 101 [61] TSKinFace [158]
3		TSKinFace [158]				TSKinFace [158]			TSKinFace [158]
4		family 101 [61] CornellKin [62] UB KinFace [218] KinFaceW-I&II [129] TALKIN [216] KFVW [226]	CornellKin [62] UB KinFace [218] family 101 [61]	KFVW [226] TALKIN [216]	KinFaceW-I&II [129]		TALKIN [216] KFVW [226]	TALKIN [216]	
7	UvA-smile [51]	KIVI [101]		KIVI [101]	UvA-smile [51]		UvA-smile [51] KIVI [101]		KIVI [101]
11		FIW [166] FIW-MM [163]				FIW [166] FIW-MM [163]			FIW [166] FIW-MM [163]

larger datasets are proposed to support data-driven methods. Based on the kin-type numbers, existing datasets can be divided into three categories: 4-types, 7-types, and 11-types (non-kin is not considered). The development of public kinship datasets is shown in Figure 6. As depicted, the blue box represents image-based datasets and green boxes correspond to video-based datasets. It is shown that, in the early days, kinship datasets are mainly image-based. Recently, video-related datasets are collected, and their labels are becoming more diversified. Table 1 lists the similarities and differences between datasets in a checkerboard manner. The darker the block is, the more similar the datasets are. Most datasets contain four kin-types, and most of these images are unconstrained. The number of images is usually less than 1000.

### 2.3.1 Kinship Datasets: 4-types

#### Image-based dataset

**CornellKin (2010)** [62]: CornellKin<sup>1</sup> is the first widely used public dataset collected in 2010, consisting of 150 pairs of public persons and celebrities with family information. The images are collected through a controlled online search, with frontal pose and neutral facial expressions. These datasets can be divided into four categories: Father-Son (F-S, 40%), Father-Daughter (F-D, 22%), Mother-Son (M-S, 13%), Mother-Daughter (M-D, 26%) with different race (around 50% Caucasians, 40% Asians, 7% African Americans, and 3% others), gender, and age.

**UB KinFace (2011)** [218, 220]: Different from CornellKin, UB KinFace<sup>2</sup> contains three images for each positive set with 270 images collected in total and separated into

<sup>1</sup> <http://chenlab.ece.cornell.edu/projects/KinshipVerification/>

<sup>2</sup> <http://www1.ece.neu.edu/~yunfu/research/Kinface/Kinface.htm>

90 groups. Each group contains three types of images: child, young parent, and old parent. This dataset is updated into the so-called UB KinFace Ver2.0 in which groups are extended from 90 to 200. For UB KinFace Ver2.0, the influence of ethnicity is considered. There are four kin-types (F-S, F-D, M-S, M-D). To our knowledge, UB KinFace is the first database collecting children, young parents, and old parents for kinship verification. However, Yan *et al.* [229] show that there is a large imbalance in the dataset because nearly 80% of UB KinFace are father-son relationships.

**Family101 (2013)** [61]: Family101<sup>3</sup> is collected based on the family trees. It contains 101 different family trees with 206 nuclear families. Each family tree contains 1 to 7 families. It consists of renowned (public) families. Each family contains 3 to 9 family members. This dataset contains 72% Caucasians, 23% Asians, and 5% African Americans with different gender or age. There are 607 individuals and 14816 images in total. Family101 is organized by a family structure providing a more structure-related task for kinship recognition.

**KinFaceW-I & KinFaceW-II (2014)** [129, 131]: The aforementioned public datasets are relatively small. Lu *et al.* [129, 131] collect KinFaceW-I and KinFaceW-II<sup>4</sup> datasets through an online search. These two datasets are slightly different during data collection. In KinFaceW-I, kinship pairs are collected from different pictures. In KinFaceW-II, all pairs are obtained from the same photo. These photos are unconstrained in terms of pose, lighting, background, expression, age, ethnicity, and partial occlusion [131]. There are four types of kinship relations for these two datasets. In KinFaceW-I, there are 156 pairs of F-S, 134 pairs of F-D, 116 pairs of M-D, and 127 pairs of M-S. Meanwhile, in KinFaceW-II, there are 250 pairs of pictures for each kinship relation.

**TSKinFace (2015)** [158]: Most of the publicly available datasets are bi-subject *i.e.*, based on a pair of images. In contrast, TSKinFace<sup>5</sup> is a tri-subject kinship database for kinship verification in a one-versus-two manner. There are two types in TSKinFace: Father-Mother-Son (FM-S) and Father-Mother-Daughter (FM-D). The FM-S contains 513 groups of tri-subjects. 343 groups of them are Asian, and 170 groups are non-Asian. FM-D contains 502 groups. There are 331 Asian groups and 171 non-Asian groups.

#### *Video-based datasets*

As opposed to still images, videos contain face dynamics including changes in head movements, expressions, and illumination conditions [226]).

**KFVW (2018)** [226]: The Kinship Face Videos in the Wild (KFVW) dataset contains 418 pairs of video clips from TV shows on the Internet. Each clip contains 100 to 500 frames. Videos are unconstrained in pose, lighting, background, occlusion, expression, makeup, and age. The average size of the video frames is about  $900 \times 500$  pixels. Similar

3 <http://chenlab.ece.cornell.edu/projects/KinshipClassification/index.html>

4 <http://www.kinfacew.com/>

5 [http://parnec.nuaa.edu.cn/\\_upload/tpl/02/db/731/template731/pages/xtan/TSKinFace.html](http://parnec.nuaa.edu.cn/_upload/tpl/02/db/731/template731/pages/xtan/TSKinFace.html)

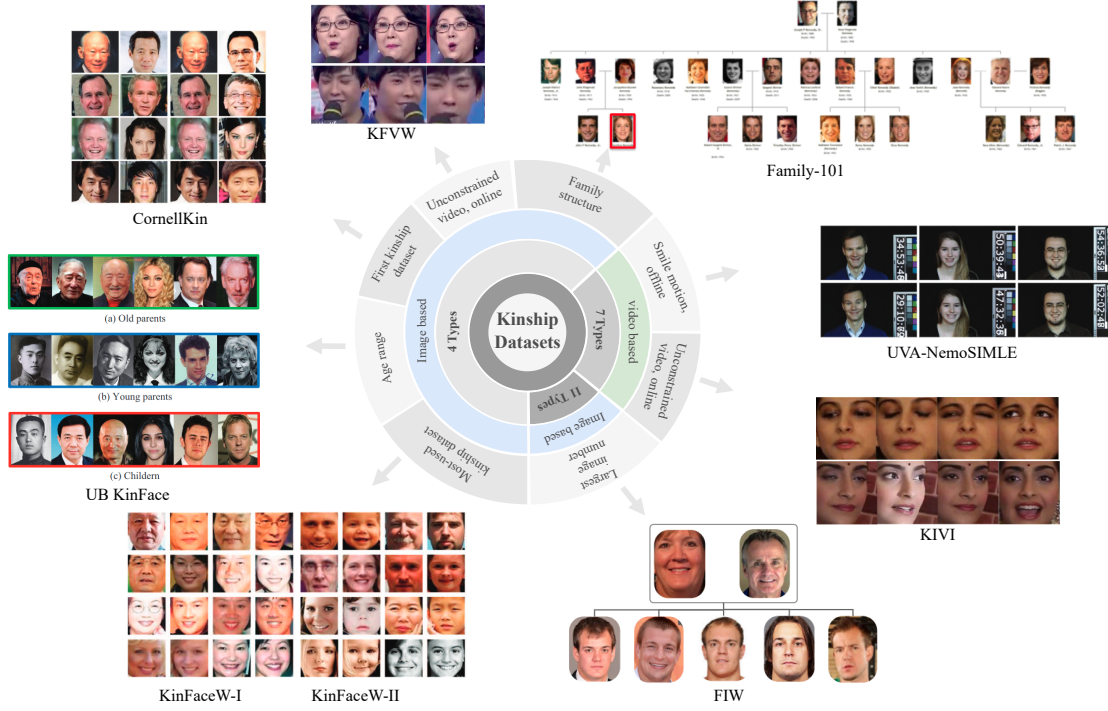


Figure 7: Representative public kinship datasets.

to other datasets, there are four types of kinship relations in KFWW: F-S, F-D, M-S, and M-D, with 107, 101, 100, and 110 pairs of videos, respectively.

**TALKIN (2019)** [216]: Wu *et al.* [216] collect a multi-model kinship dataset called TALKing KINship (TALKIN). This dataset is collected from YouTube with a prepared list of celebrities and family TV shows. It consists of both visual and audio information. After collection, the data is cropped and resized into  $224 \times 224$  resolution. There are four kin relations in the dataset: F-S, F-D, M-S, and M-D. Each relation contains 100 pairs of videos. The length of the videos ranges from 4.032 seconds to 15 seconds.

### 2.3.2 Kinship Datasets: 7-types

#### Image-based datasets

**WVU (2016)** [99]: WVU is collected by Kohli *et al.* in 2017 and contains 113 pairs. The statistics for each kin-type are 22 (B-B), 9 (B-S), 13 (S-S), 14 (F-D), 32 (F-S), 13 (M-D), and 8 (M-S).

**IIITD (2012)** [98]: IIITD contains 272 pairs of celebrities on the Internet. It consists of four ethnicities: Afro-American, American, Indian, and Asian. The numbers of each kin-type are 42 (B-B), 49 (B-S), 55 (S-S), 33 (F-D), 2 (F-S), 26 (M-D), and 52 (M-S).

#### Video-based datasets

**UvA-NEMO Smile (2013)** [49, 51]: UvA-Nemo smile dataset contains 1240 videos of 400 subjects with a resolution of  $1920 \times 1080$  at 50 *fps* rate. The dynamics of

spontaneous and posed smiles of each subject are recorded. All videos are constrained *i.e.*, keeping the same viewing angle and background. There are 95 kin relations for the 152 subjects. Videos contain two types of smiles: 228 pairs of spontaneous and 287 pairs of posed smiles. There are seven kin relationships: S-S, B-B, S-B, M-D, M-S, F-D, and F-S.

**KIVI(2019)** [101]: KIVI<sup>6</sup> is collected from the Internet to include realistic in-the-wild variations. It contains 503 videos of individuals from 211 families. There are 355 positive kin pairs. The videos' duration is around 18.78 seconds, with a frame rate of 26.79 frames per second (fps). The total number of still frames in the database is over 250,000 [101].

### 2.3.3 Kinship Datasets: 11-types

#### *Image-based datasets*

**FIW (2016)** [166] [168] [198] [170]: Over time, datasets with larger capacities and more kinship types are provided. The Families In the Wild (FIW<sup>7</sup>) dataset is collected from the Internet. Over 10000 family photos of 1000 families are labelled. There are 11193 unconstrained family photos of 1000 families. On average, there are ten images for each family. Later, Robinson *et al.* [167] [169] extended FIW. In [167], over 13000 family photos of 1000 families are collected. The number of pairs increased from 418000 to 656954. In [169], existing labels are used for each family as side information to add more data to under-represented families.

#### *Video-based datasets*

**FIW-MM (2020)**: FIW with multimedia (FIW-MM) dataset [163] is a dataset proposed by Robinson *et al.*. It is an extended version of the FIW dataset. FIW-MM extended the existing paired faces of FIW via an automated labelling pipeline. Multimedia (MM) data (*i.e.*, video, audio, and text captions) is collected.

### 2.3.4 Others

In addition to the datasets mentioned above, datasets used for other applications also contain kinship information. The Family Face Database (FF-Database) [246] is used for the face prediction of children. It consists of 7488 parent and 8558 child faces with  $128 \times 128$  resolution. Six facial attributes are labelled: expression, gender, age, glasses, moustache, and skin colour. The People in Social Context (PISC) [110] dataset is collected for the task of social relation recognition. It consists of common social relationships, including commercial, couples, family, friends, *etc.*. The People in Photo Album (PIPA) dataset [183] is collected from Flickr photo albums and can be used for both person recognition and social relation recognition. 16 finer relationships are labelled,

<sup>6</sup> <http://iab-rubric.org/resources/KIVI.html>

<sup>7</sup> <https://web.northeastern.edu/smilelab/fiw/>



including the grain kinship relationships, such as father-child. Although these datasets are created for other tasks, they can also be used for kinship verification.

### 2.3.5 Discussion

In Figure 6 and Table 1, it is shown that image-based kinship datasets are well-developed for image-based kinship verification. In contrast, there is still a demand for video-based kinship datasets. According to Table 1, most of the datasets are collected in unconstrained settings, causing many external interference factors, and making it difficult to study kinship verification systematically.

Several kinship datasets can be used to study specific kinship problems. For example, TSKinFace can be used for the tri-subject kinship verification task. UB KinFace is suitable for kinship verification of elderly people, and TALKIN for multi-modal and sound-based kinship verification. In contrast to specific kinship problems, general-purpose kinship and generic datasets are required. To this end, we collected the Nemo-kinship dataset for the purpose of child-adult kinship verification. This dataset is discussed in Section 5.

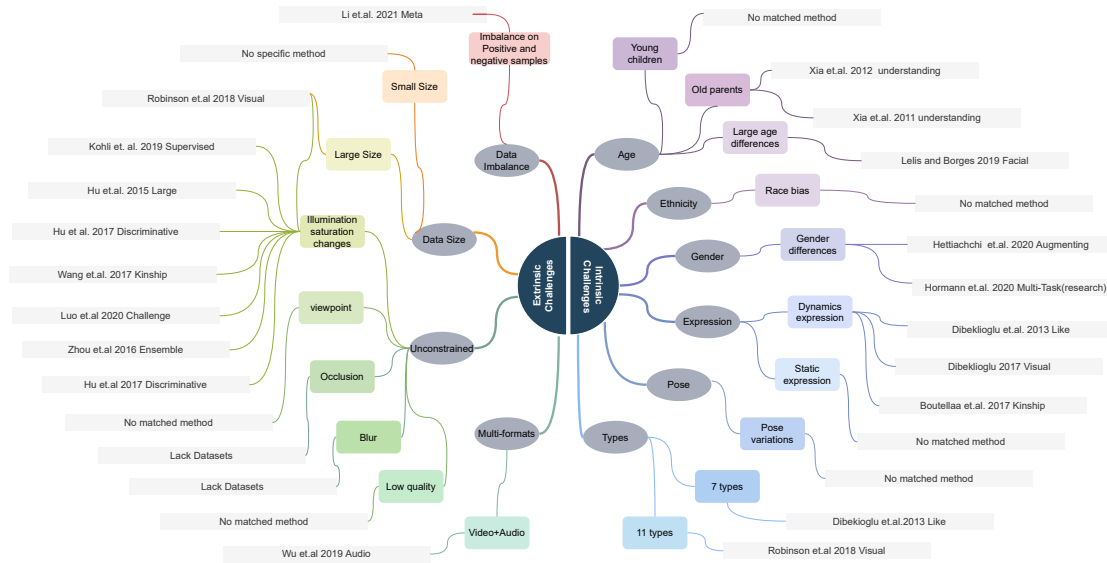


Figure 8: Challenges and related methods.

## 2.4 REPRESENTATIVE METHODS

Figure 8 shows the challenges summarized for kinship verification and corresponding approaches. There are six internal sub-challenges *i.e.*, age, race, gender, facial expression, posture, and kin-type. There are four for extrinsic *i.e.*, data imbalance, data size, unconstrained, and multi-modal. We select and list the corresponding approach. Many challenges have their corresponding methods, but some of them lack a specific solution. For example, there are currently no kinship verification methods proposed to deal with racial bias or low-quality facial images.

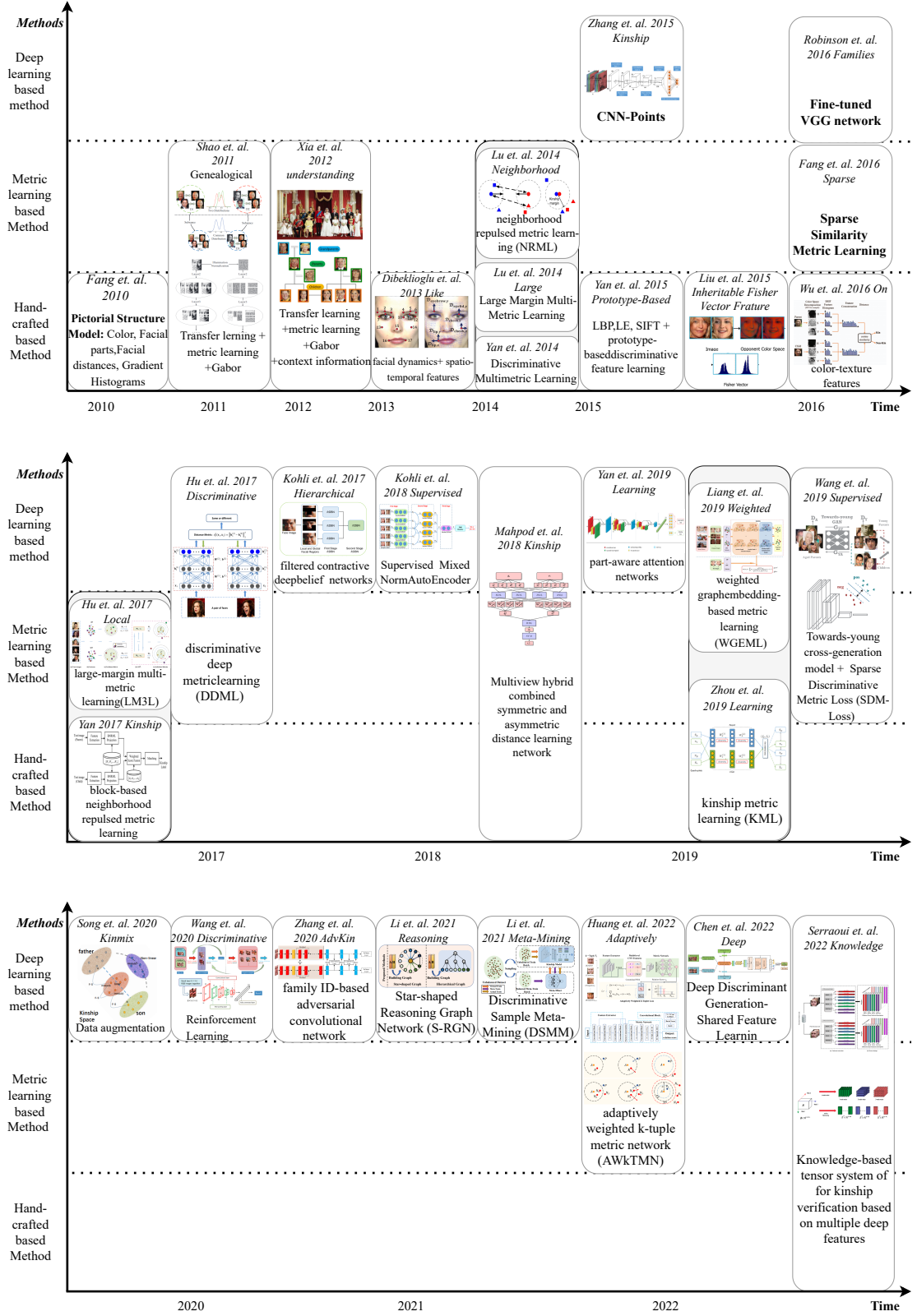


Figure 9: Milestones of kinship verification methods.

Figure 9 shows the development of existing methods. According to the type of input, kinship verification methods can be divided into image-based and video-based methods. Among each of them, we divide the methods into three categories according to their feature representation: (1) hand-crafted feature-based, (2) metric learning-based, and (3) deep learning-based. The hand-crafted feature-based category includes traditional hand-crafted descriptors. The extracted facial features are used by standard discriminators such as KNN and SVM. The metric-learning-based category mainly focuses on projecting latent features onto more prominent spaces. The goal of these methods is to decrease the intra-class distance of the projected features and to increase the inter-class distance. The third category is based on deep learning, such as CNNs, GANs, GCNs, and auto-encoders.

#### 2.4.1 Image-based

##### *Handcrafted feature descriptors*

The first kinship verification method is proposed by Fang *et al.* [62]. The method uses 22 hand-crafted (facial) features to represent the geological information between parents and children. These features are low-level features such as color, facial geometry, and texture. Then, K-Nearest-Neighbors and SVMs are trained based on these features. The top 14 factors are selected based on the classification accuracy. It shows that most of the informative parts are around the eyes. Since these features correspond to local parts, global features are also included. Later, Fang *et al.* [61] use the dense SIFT (dSIFT) descriptor for kinship verification. After this first publication, different hand-crafted feature extracting methods are proposed [15, 54, 107, 130, 158, 159, 210, 228, 230, 249]. Low-level features such as HOG [207], LBP [3], LPQ are used for kinship verification.

An overview of hand-crafted descriptors is as follows:

**SIFT:** Scale-invariant feature transform (SIFT) [126]. A series of kinship verification methods [130, 158, 159, 228, 230] use SIFT to extract kinship features.

**LBP:** Local binary patterns (LBP) are used for face recognition [2, 227]. There are different variants of LBP: Three-patch LBP (TPLBP) [212], CLBP and ILBP [91] *etc.*. For kinship verification methods, Boutellaa *et al.* [18] use LBP and LBPTOP features. Lu *et al.* [130] exploit TPLBP features. Zhou *et al.* [249], Dornaika *et al.* [54] and Wei *et al.* [210] use LBP as one of the feature descriptors to verify the kinship.

**Gabor** [1, 133]: Zhou *et al.* [251] utilize a Gabor wavelet and propose a Gabor-based gradient orientation pyramid feature for kinship verification. Xia *et al.* [221], and Shao *et al.* [176] partition the face into regions in multiple layers and then compute Gabor filters in each region to extract genetic-invariant features.

According to Yan and Lu [227], due to the large variations of faces caused by varying imaging conditions, low-level feature descriptors such as LBP and SIFT may fall short. Therefore, new approaches are proposed, such as the spatial pyramid learning-based (SPLE) feature descriptor [249] to automatically exploit both local appearance and global

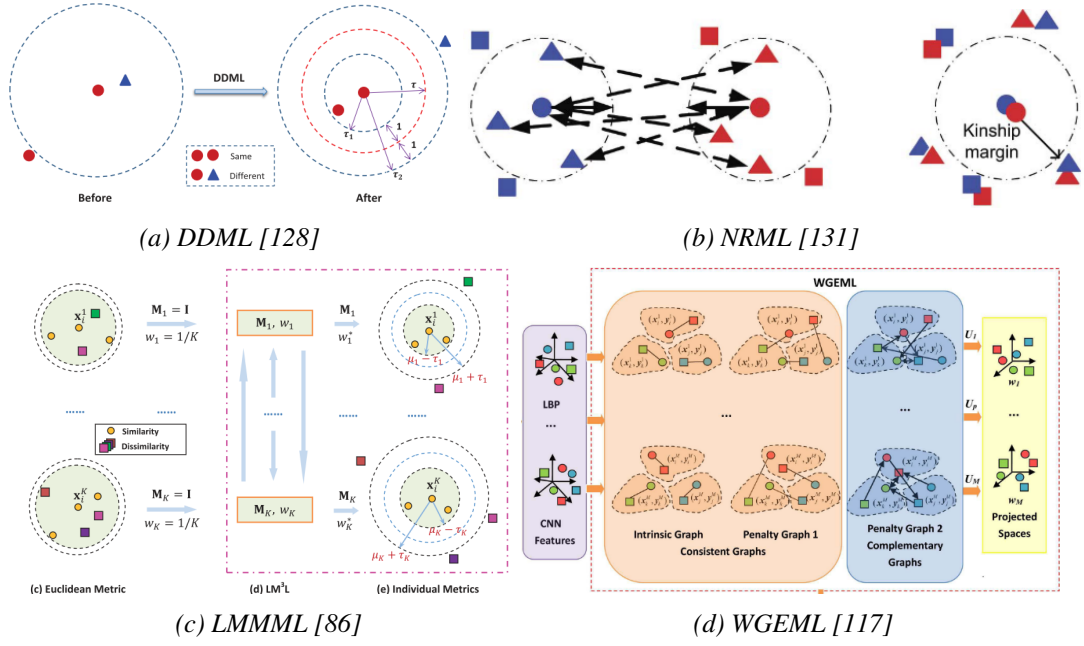


Figure 10: Illustration of metric learning-based methods, cited from [86, 117, 128, 131].

spatial information. The SPLE obtains improved results compared to PCA, LBP, HOG, and LE [13]. An extension of the method is provided using a new Gabor-based Gradient Orientation Pyramid (GGOP) [251].

Other methods focus on combining feature detectors such as Alirezazadeh *et al.* [4] targeting a combination of local and global hand-crafted features resulting in improved results (81.3% and 86.15% on dataset KinFaceW-I and KinFaceW-II, respectively). Later, Boutellaa *et al.* [18] use spatio-temporal features based on a combination of hand-crafted LBP, LPQ, and BSIF, and deep learning features.

### Metric learning

Different from handcrafted-feature-based methods, metric learning-based methods focus on the similarity measurement itself *i.e.*, decreasing the intra-class and increasing the inter-class distance of the facial features (samples) [70, 157, 227, 254]. It learns a distance metric to measure the similarity between samples [128]. Metric learning can be divided into two categories [131, 227]: unsupervised and supervised. Unsupervised methods use principal component analysis (PCA) [188], linear discriminant analysis (LDA) [12], and Laplacian eigenmaps (LE) [13]. For supervised methods, the Mahalanobis distance metric is often utilized. The distance function of an image pair  $\mathcal{P} = \{(X_i, Y_i) | i = 1, 2, \dots, N\}$  is described by  $d(x_i, y_j) = \sqrt{(x_i - y_j)^T M (x_i - y_j)}$ , where  $M$  is a  $m \times m$  square matrix, and  $1 \leq i, j \leq N$ .  $M$  is further decomposed into  $W^T W$ , which converts the Mahalanobis distance to  $d(x_i, y_j) = \|Wx_i - Wy_j\|_2$ .  $x_i$  and  $y_j$  are feature vectors of  $X_i, Y_j$  extracted from  $g(\cdot)$ . The target is transformed from a learning distance metric  $M$  to seeking a linear transformation  $W$  which projects the input  $x_i, y_j$  into a more suitable subspace. Ensemble metric learning [179], neighborhood repulsed metric learning (NRML) [131], large

margin multi-metric metric learning (LMMML) [86], and discriminative multi-metric learning (DMML) [52, 198] are representative methods.

**NRML** [131]: Neighborhood repulsed metric learning (NRML) is proposed by Lu *et al.* [129]. An extension is provided by [131]. Lu *et al.* propose NRML to ensure that the intra-class samples are close to each other and repulse the inter-class samples as far as possible. Previous metric learning methods consider the samples equally, whereas NRML determines more informative samples as follows:

$$\begin{aligned}
\max_{\mathbf{M}} J(\mathbf{M}) &= J_1(\mathbf{M}) + J_2(\mathbf{M}) - J_3(\mathbf{M}) \\
&= \frac{1}{Nk} \sum_{i=1}^N \sum_{t_1=1}^k d^2(\mathbf{x}_i, \mathbf{y}_{it_1}) \\
&\quad + \frac{1}{Nk} \sum_{i=1}^N \sum_{t_2=1}^k d^2(\mathbf{x}_{it_2}, \mathbf{y}_i) - \frac{1}{N} \sum_{i=1}^N d^2(\mathbf{x}_i, \mathbf{y}_i), \\
&= \frac{1}{Nk} \sum_{i=1}^N \sum_{t_2=1}^k (\mathbf{x}_{it_2} - \mathbf{y}_i)^T \mathbf{A} (\mathbf{x}_{it_2} - \mathbf{y}_i) \\
&\quad - \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \mathbf{y}_i)^T \mathbf{A} (\mathbf{x}_i - \mathbf{y}_i)
\end{aligned} \tag{3}$$

where  $\mathbf{y}_{it_1}$  represents the  $t_1$ th k-nearest neighbor of  $\mathbf{y}_i$ , and  $\mathbf{x}_{it_2}$  denotes the  $t_2$ th k-nearest neighbor of  $\mathbf{x}_i$ . The optimization function is solved by determining k-nearest neighbors of  $\mathbf{x}_i$  and  $\mathbf{y}_i$  based on the Euclidean metric and then solve  $d$  sequentially.

**DMML** [229]: Yan *et al.* [229] propose discriminative multi-metric learning (DMML). DMML aims to extract multiple features to exploit more complementary information by jointly learning multiple distance metrics. Unlike NRML, DMML tries to maximize the probability instead of directly minimizing the intra-class distance and maximizing the inter-class distance. In this method, each pair of positive samples has the highest probability of having a shorter distance than the most similar negative sample. In addition, the correlation between different features is also maximized. The DMML method can be formulated as a constrained optimization problem as follows:

$$\begin{aligned}
\min_{\mathbf{W}_1, \dots, \mathbf{W}_K, \alpha} J &= \sum_{k=1}^K \alpha_k f_k(\mathbf{W}_k) \\
&\quad + \lambda \sum_{\substack{k_1, k_2=1 \\ k_1 \neq k_2}}^K \sum_{i=1}^N \|\mathbf{W}_{k_1}^T \mathbf{x}_i^{k_1} - \mathbf{W}_{k_2}^T \mathbf{x}_i^{k_2}\|_F^2,
\end{aligned} \tag{4}$$

where  $f_k(\mathbf{W}_k) = \Pi_{O_1^k} \log(1 + \exp(\|\mathbf{W}_k^T \mathbf{x}_{ik}^p\|^2 - \|\mathbf{W}_k^T \mathbf{x}_{ik}^n\|^2))$  and  $\mathbf{x}_{ik}^p = \mathbf{x}_i^k - \mathbf{y}_i^k$ ,  $\mathbf{x}_{ik}^n = \mathbf{x}_i^k - \mathbf{y}_j^k$ , where  $\mathbf{x}_i^k$  and  $\mathbf{y}_j^k$  represent  $k$ th feature of  $\mathbf{X}_i$  and  $\mathbf{Y}_i$ . The first term augments the probability that a negative pair distance is larger than the positive pair distance. The second term ensures that different features reach as much complementary information as possible. Since the equation has no closed-form solution, Yan *et al.* firstly initialize  $\mathbf{W}_k$  and  $\alpha$ , and update  $\mathbf{W}_k$  sequentially by using the gradient descent method, where  $\alpha$  is updated accordingly.

Table 2: Different methods.

Year	Method	Author	Dataset	Data type	Input	Evaluation	key Idea
2010	FPFD+ SVM	Fang <i>et al.</i> [62]	Cornell KinFace	images	4 kin relations	70.67	Handcrafted feature
2011	transfer learning	Xia <i>et al.</i> [218]	UB KinFace	images	3 subsets	60	Transfer subspace learning
2011	SPLE +SVM	Zhou <i>et al.</i> [249]	private Dataset	images	4 kin relations	67.75	Local appearance + global spatial information;automatically
2012	GGOP	Zhou <i>et al.</i> [251]	private Dataset	images	4 kin relations	69.75	Gabor wavelet+ Gradient Orientation Pyramid feature +SVM
2012	Gabor+ TSL	Xia <i>et al.</i> [218] [221]	UB KinFace v2	images	3 subsets	56.5	Stabilize the target distribution
2012	TSL+ age+ position	Xia <i>et al.</i> [221]	FamilyFace	images	family images(4ks)	79.66	Make use of additional information
2013	dynamic+ CLBP-TOP	Dibeklioglu <i>et al.</i> [49]	UvA-NEMO Smile	video	7 kin relations	67.11	Dynamic+ spatial temporal information
2014	NRML	Lu <i>et al.</i> [129, 131]	KinFace-W-I KinFace-W-II	images	4 kin relations 4 kin relations	69.9 76.5	Repulse neighborhood samples by using KNN +metric learning
2013	Graph based	Guo <i>et al.</i> [78]	Sibling-Face Database	images	7 kin relation	69.25	Take advantages of graph information
2014	DMMML	Yan <i>et al.</i> [228]	KinFace-W-I KinFace-W-II	images	4 kin relations 4 kin relations	72.5 78.25	Make use of multi descriptors Metric Learnin
2014	LMMML	Hu <i>et al.</i> [86]	KinFaceII	images	4 kin relations	81.28	Utilize large margin multi-metric learning and add threshold
2015	RSBM	Qin <i>et al.</i> [158]	TSKinFace	images	two family type	85.4	Family info + symmetric bilinear model
2015	PDFL	Yan <i>et al.</i> [230]	KinFace-W-I	images	4 kin relation	70.1	Mid-level features by discriminative learning
			KinFace-W-II	images	4 kin relation	77	
			Cornell KinFace	images	4 kin relation	71.9	
			UB KinFace	images	4 kin relation	67.3	
2015	CNN-Points	Zhang <i>et al.</i> [240]	KinFace-W-I	images	4 kin relation	77.5	CNN + key points structure
			KinFace-W-II	images	4 kin relation	88.4	
2015	LIRIS	Lu <i>et al.</i> [127]	KinFace-W-I	images	4 kin relations	86.3	Multi feature descriptor selection + feature selection + SVM
			KinFace-W-II	images	4 kin relation	83.1	
2015	Genetic Measure	Kou <i>et al.</i> [103]	KinFace-W-I	images	4 kin relation	65.7	Use sparse similarity measure
			KinFace-W-II	images	4 kin relation	74.8	
2016	SSSL	Xu <i>et al.</i> [223]	KinFace-W-I	images	4 kin relation	77.2	Jointly learn multiple sparse bilinear similarity models (structured similarity Fusion)
			KinFace-W-II	images	4 kin relation	76	
2016	Fine-tune CNNs	Robinson <i>et al.</i> [166]	FIW	images	11 kin relations	71	fine-tune VGG-Face CNN
2016	Deep+ shallow	Boutellaa <i>et al.</i> [18]	UvA-NEMO Smile	video	7 kin relation	90.98	Combine Spatio-Temporal Texture Features and deep feature
2016	LC-FS (HDLBPH)	Zhang <i>et al.</i> [239]	TSKinFace	images	three subsets	89.7	Modeling genetic transferring by two parents information
2017	MLCSL	Chen <i>et al.</i> [29]	KinFace-W-I	image	4kin relations	83.28	Multi-linear coherent space learning with multi-scale features
			KinFace-W-II	image	4kin relations	84.3	
2017	DML	Wang <i>et al.</i> [198]	FIW	images	9 kin relations	68.79	Metric learning+ Denoising Auto-encoder
2017	DDMML	Lu <i>et al.</i> [128]	KinFace-W-I	images	4kin relations	83.5	Hierarchical representation+ filtered contractive deep belief networks
			KinFace-W-II	images	4kin relations	84.5	
			TSKinFace	images		85.3	
2017	visual transformation	Dibeklioglu <i>et al.</i> [48]	UvA-NEMO Smile	videos	7 kin relations	93.65	Deep contrastive learning architecture
2017	KVRL-fcDBN	Kohli <i>et al.</i> [99]	KinFace-W-I	images	4kin relations	96.1	Hierarchical representation+ filtered contractive deep belief networks
2017	KinNet	Robinson <i>et al.</i> [170]	KinFace-W-II	images	4kin relations	96.2	
2017	video-based DML	Yan <i>et al.</i> [226]	subset FIW	images	7 kin relations	74.86	Fine-to-coarse deep metric learning with triplet loss
2017	CFT	Duan <i>et al.</i> [56]	KinFace-W-I	images	4 kin relations	77.4	Coarse-to-Fine Transfer Learning +NRML
			KinFace-W-II	images	4 kin relations	79.3	
			UB KinFace	images	4 kin relations	72.3	
			Cornell KinFace	images	4 kin relations	78.6	
2017	LLMMML	Hu <i>et al.</i> [85]	KinFace-W-I	images	4 kin relations	80	Jointly learns multiple distance metrics
2018	SDM-Loss	Wang <i>et al.</i> [197]	FIW	images	11 kin relations	69.47	GAN +sparse discriminative metric loss
2018	SphereFace	Robinson <i>et al.</i> [168]	FIW	images	11 kin relations	69.18	Fine-tuned CNNs +angular softmax loss
2019	attention network	Yan <i>et al.</i> [232]	KinFace-W-I	images	4 kin relations	82.6	Multi input + attention network
			KinFace-W-II	images	4 kin relations	92	
2020	GKR network	Li <i>et al.</i> [114]	KinFace-W-I	images	4 kin relations	79.2	Kinship Relational Graph + MLP
			KinFace-W-II	images	4 kin relations	90.6	
2020	KinMix	Song <i>et al.</i> [180]	KinFace-W-I	images	4 kin relations	78.5	Convolutional neural network + Linear generation in feature space
			KinFace-W-II	images	4 kin relations	89.7	
2020	NESN-KVN	Wang <i>et al.</i> [199]	KinFace-W-I	images	4 kin relations	78.6	Deep convolutional networks + reinforcement learning
			KinFace-W-II	images	4 kin relations	89.0	
2021	DSMM	Li <i>et al.</i> [113]	KinFace-W-I	images	4 kin relations	82.4	Meta-miner network + CNN
			KinFace-W-II	images	4 kin relations	93.6	
2021	AdvKin	Zhang <i>et al.</i> [241]	KinFace-W-I	images	4 kin relations	79.6	Family ID-based adversarial convolutional network+ self-adversarial mechanism
			KinFace-W-II	images	4 kin relations	89.9	
			UB KinFace	images	4 kin relations	75	
			Cornell KinFace	images	4 kin relations	81.4	
2021	Relational Network	Yan <i>et al.</i> [231]	KinFace-W-I	images	4 kin relations	85.6	Deep relational network + multi-scale features
			KinFace-W-II	images	4 kin relations	88.8	
2022	D2GFL	Chen <i>et al.</i> [30]	KinFace-W-I	images	4 kin relations	83.1	Generation-shared feature + CNN
			KinFace-W-II	images	4 kin relations	88.7	
			Cornell KinFace	images	4 kin relations	82.9	
			TSKinFace	images	4 kin relations	91.3	
2022	AWK-TMN	Huang <i>et al.</i> [89]	KinFace-W-I	images	4 kin relations	80.4	Metric Learning + adaptively weighted scheme + multiple levels of convolutional features
			KinFace-W-II	images	4 kin relations	91.6	
2022	TXQEDA+WCCN	Serraoui <i>et al.</i> [174]	KinFace-W-I	images	4 kin relations	91.11	Multi-view deep feature extraction + Metric learning
			KinFace-W-II	images	4 kin relations	90.30	
			Cornell KinFace	images	4 kin relations	90.68	
			TSKinFace	images	4 kin relations	93.77	

**DML** [198] : Discriminative metric learning uses a linear projection [52] defined by:

$$\min_{\mathbf{W}\mathbf{W}^T=\mathbf{I}} \frac{\text{tr}(\mathbf{W}\mathbf{F}\mathbf{L}_w\mathbf{F}^T\mathbf{W}^T)}{\text{tr}(\mathbf{W}\mathbf{F}\mathbf{L}_b\mathbf{F}^T\mathbf{W}^T)}, \quad (5)$$

where  $\mathbf{F} = [\mathbf{x}_i, \dots, \mathbf{x}_n, \mathbf{y}_i, \dots, \mathbf{y}_n]$  denotes the training data, and  $\mathbf{L}_w$  and  $\mathbf{L}_b$  are Laplacian matrices. Wang *et al.* [198] use denoising auto-encoder-based robust metric learning by combining denoising auto-encoding (DAE) and metric learning. The projection matrix is constrained simultaneously by both DAE and metric learning to obtain a nonlinear transformation. The loss of DML is given by:

$$\mathcal{L} = \min_{\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2} \frac{1}{2} \|\mathbf{F} - \hat{\mathbf{F}}\|_F^2 + \frac{\lambda}{2} \text{tr} \left( \frac{\mathbf{H}\mathbf{L}_w\mathbf{H}^T}{\mathbf{H}\mathbf{L}_b\mathbf{H}^T} \right), \quad (6)$$

where  $\mathbf{H} = \sigma(\mathbf{W}_1 \mathbf{F} + \mathbf{B}_1)$ ,  $\hat{\mathbf{F}} = \sigma(\mathbf{W}_2 \mathbf{H} + \mathbf{B}_2)$ , and  $\mathbf{B}_1$  and  $\mathbf{B}_2$  are the offset matrices. The projection matrix  $\mathbf{W}$  is used as an encoded hidden layer. The DML encodes the feature non-linearly while maximizing the inter-class distance and minimizing the intra-class distances.

**DDML** [128]: Deeper non-linear representations are preferred, since linear transformations are shallow and may not be powerful enough. Similar to DML, discriminative deep metric learning (DDML) uses a deep neural network to learn a set of hierarchical nonlinear transformations to project pairs into an optimized feature space. Hu *et al.* [128] propose a deep neural network  $f(\cdot)$  to generate representations of sample pairs. Sample pairs are fed into the network non-linearly. The Euclidean distance of these representations is defined by  $d_f^2(\mathbf{X}_i, \mathbf{X}_j) = \|f(\mathbf{X}_i) - f(\mathbf{X}_j)\|_2^2$ . A margin framework is used to separate positive and negative pairs. As illustrated in Figure 10a, a threshold  $\tau$  ( $\tau > 1$ ) is used to enforce the distance of a positive pair ( $l_{ij} = 1$ ) to be smaller than  $\tau$  and the distance of a negative pair ( $l_{ij} = -1$ ) to be larger than  $\tau$ . The optimization function is defined by:

$$\begin{aligned} \arg \min_f J &= J_1 + J_2 \\ &= \frac{1}{2} \sum_{i,j} g(1 - \ell_{ij}(\tau - d_f^2(\mathbf{X}_i, \mathbf{X}_j))) \\ &\quad + \frac{\lambda}{2} \sum_{m=1}^M (\|\mathbf{w}^{(m)}\|_F^2 + \|\mathbf{b}^{(m)}\|_2^2) \end{aligned} \quad (7)$$

where  $g(z) = \frac{1}{\beta} \log(1 + \exp(\beta z))$  is a logistic loss function and  $\beta$  is a sharpness parameter.  $\|\cdot\|_F$  represents the Frobenius norm of the matrix.  $\lambda$  is a regularization parameter.

In conclusion, the combination of deep learning with the discriminative ability of metric learning is one of the promising directions of current methods.

### Deep learning

Previous studies [21, 88, 184] show that deeper layers extract higher-level features effectively.

**CNNs:** Zhang *et al.* [240] use, for the first time, a deep learning model. The proposed convolutional neural network consists of three convolutional layers and one fully connected layer. To use local information, images are cropped into different patches based on their facial landmarks. Then, the aligned patches are fed into the matched sub-models. A significant improvement is obtained compared to earlier methods [129–131, 228]. From that moment on, different CNN-based methods [7, 30, 33, 37, 54, 83, 106, 119, 142, 175, 180, 195, 199, 215, 232, 235, 238, 241] are proposed.

In contrast to Zhang *et al.*, Yan *et al.* [232] focus on attention mechanisms. They design a part-aware attention network to extract local facial information. Moreover, key point masks are added to the input images for a better guidance. The architecture is illustrated in Figure 11. Furthermore, Chen *et al.* [30] propose a two-stream convolutional neural network to learn parent-specific and child-specific features. Yan *et al.* [231] suggest



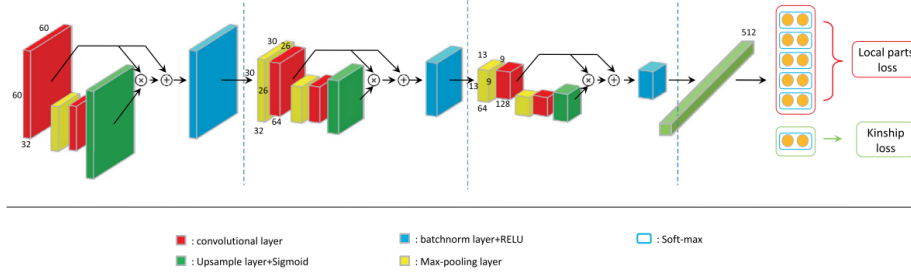


Figure 11: The structure of the attention network, cited from Yan et al. [232].

a deep relational network, utilizing multi-scale features from different convolutional layers. Wang *et al.* [199] propose a reinforcement learning-based network. They design a negative example sampling network to select more suitable samples for learning discriminative features.

In addition to kinship information, other face-related information can be used. Zhang *et al.* [241] propose a two-stream adversarial convolutional network (AdvKin) model based on family ID information. A self-adversarial strategy is exploited to reduce feature distribution discrepancy. Hormann *et al.* [83] focus on opposite-gender pairs and propose a comparator framework with kinship relation information. Song *et al.* [180] propose a KinMix method to generate positive samples in the feature space. They assume that the linearly combined kinship features yield similar clustering.

Table 3: Checkerboard of different datasets and corresponding methods.

kin-types	constrained	unconstrained	pairs <300	300 <pairs<500	500 <pairs<1000	1000>pairs	images	videos	audio	family structure
2 types		family 101 [61] TSKinFace [158]			family 101 [61]  Fang2013 [61]	TSKinFace [158]	family 101 [61] TSKinFace [158]			family 101 [61] TSKinFace [158]
		Fang2013 [61], RSBM [158], LC-FS [239], DDMML [128]				RSBM [158], LC-FS [239], DDMML [128]	Fang2013 [61], RSBM [158], LC-FS [239], DDMML [128]	Fang2013 [61], RSBM [158], LC-FS [239], DDMML [128]		Fang2013 [61], RSBM [158], LC-FS [239], DDMML [128]
3 types		TSKinFace [158]				TSKinFace [158]	TSKinFace [158]			TSKinFace [158]
		RSBM [158], LC-FS [239], DDMML [128]				RSBM [158], LC-FS [239], DDMML [128]	RSBM [158], LC-FS [239], DDMML [128]			RSBM [158], LC-FS [239], DDMML [128]
4 types		family 101 [61] CornellKin [62] KinFaceW-I&II [129] TALKIN [216] KFVW [226]  NRML [131], DMML [229], LMML [86], PDFL [230], CFT [56], CNN-Points [240], LIRIS [127], Genetic measure [103], SSL [253], MLCSL [29], DDMML [128], KVRL-4cDBN [99], LLMML [86], attenNet [232], Fang2010 [62], transferlearning [218], Gabor+TSL, audio-visual [216], video-based DML [226]	CornellKin [62] UB KinFace [218] family 101 [61]  KFVW [226] TALKIN [216]  video-based DML [226], audio-visual [216]	KinFaceW-II [129]  NRML [131], DMML [229], LMML [86], PDFL [230], CFT [56], CNN-Points [240], LIRIS [127], Genetic measure [103], SSL [253], MLCSL [29], DDMML [128], KVRL-4cDBN [99], LLMML [86], attenNet [232]		Family 101 [61] CornellKin [62] UB KinFace [218] KinFaceW-II [129]  Fang2010 [62], Transfer learning [218], Gabor+TSL, NRML [131], DMML [229], LMML [86], PDFL [230], CFT [56], CNN-Points [240], LIRIS [127], Genetic measure [103], SSL [253], MLCSL [29], DDMML [128], KVRL-4cDBN [99], LLMML [86], attenNet [232]	TALKIN [216] KFVW [226]  audio-visual [216], video-based DML [226]	TALKIN [216]  audio-visual [216]		
		UvA-NEMO Smile [51] dynamic+CLBP-TOP [49], deep+shallow [18], Dibeliogh2017 [48],	KIVI [101] SMNAE [101]		KIVI [101] SMNAE [101]	UvA-NEMO Smile [51] dynamic+CLBP-TOP [49], deep+shallow [18], Dibeliogh2017 [48],		UvA-NEMO Smile [51], KIVI [101] dynamic+CLBP-TOP [49], deep+shallow [18], Dibeliogh2017 [48], SMNAE [101]		KIVI [101] SMNAE [101]
11 types		FIW [166]  FineTune CNN [168], DML [198], SDM-Loss [197], SphereFace [165]				FIW [166]  FineTune CNN [168], DML [198], SDM-Loss [197], SphereFace [165]	FIW [166]  FineTune CNN [168], DML [198], SDM-Loss [197], SphereFace [165]			FIW [166]  FineTune CNN [168], DML [198], SDM-Loss [197], SphereFace [165]

Auto-encoders, GANs, and Graph neural networks are used for kinship verification:

**Auto-encoders:** Due to the nature of preserving identical information, auto-encoders are often used to extract genetic information. Generally, the encoder obtains the latent representation by deterministic mapping  $e = f_{\theta}(x) = s(Wx + b)$ . Here,  $x$  denotes the

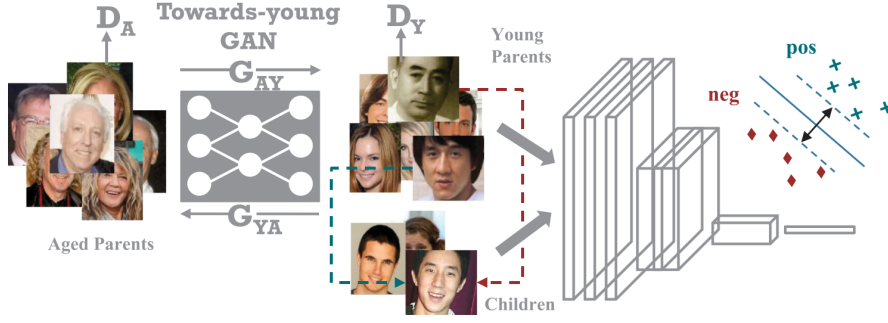


Figure 12: Illustration of a cross-generation generative kinship verification framework, cited from Wang et al. [197].

input vector. The latent representation  $\mathbf{y}$  is mapped back to reconstruct the input vector  $\hat{\mathbf{x}} = g_{\theta'}(\mathbf{e}) = s(\mathbf{W}'\mathbf{e} + \mathbf{b}')$  with the parameter  $\theta' = \{\mathbf{W}', \mathbf{b}'\}$ . The auto-encoder can be optimized by [190]:

$$\begin{aligned} \theta^*, \theta'^* &= \arg \min_{\theta, \theta'} \frac{1}{n} \sum_{i=1}^n L(\mathbf{x}^{(i)}, \hat{\mathbf{x}}^{(i)}) \\ &= \arg \min_{\theta, \theta'} \frac{1}{n} \sum_{i=1}^n L(\mathbf{x}^{(i)}, g_{\theta'}(f_{\theta}(\mathbf{x}^{(i)}))). \end{aligned} \quad (8)$$

Here  $\theta = \{\mathbf{W}, \mathbf{b}\}$ , where the loss function is  $L(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2$ . Liang *et al.* [116] use auto-encoders to learn deep relational features. Dehghan *et al.* [44] propose to use gated auto-encoders with a discriminating neural network layer. Wang *et al.* [196] propose a deep kinship verification (DKV) model and utilize metric learning methods to extract features. Firstly, they use a stacked auto-encoder network to select nonlinear low-dimension features. Then, deep kinship verification is combined with a stacked auto-encoder network and metric learning.

**GANs:** Although genetic-related information is used [44, 61, 116], these methods may fall short to deal with (test) pairs with large age differences yielding a performance drop in kinship verification accuracy [197, 218, 220]. To mitigate age and identity divergences, Wang *et al.* [197] propose a towards-young cross-generation model with a Sparse Discriminative Metric Loss (SDM-Loss). As shown in Figure 12, the aged parents are generated to a young age while keeping the same identity. Then, the image pair is extracted through a convolutional neural network constrained by SDM-loss. The derived discriminative metric minimizes the feature gap among aged parents and children, alleviating the intrinsic side effects.

**Graph neural networks:** Li *et al.* [114] propose a graph-based kinship reasoning (GKR) network that performs relational reasoning on the extracted features. The overall framework of the GKR network is shown in Figure 13. Features are extracted by the same convolutional neural network and built into a Kinship Relational Graph. A recursive

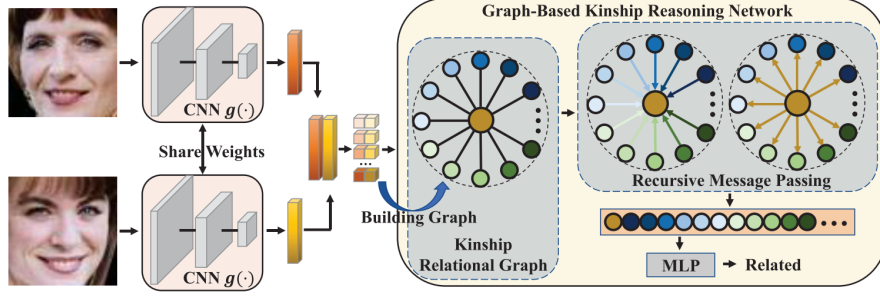


Figure 13: Framework of a graph-based kinship reasoning network, cited from Li et al. [114].

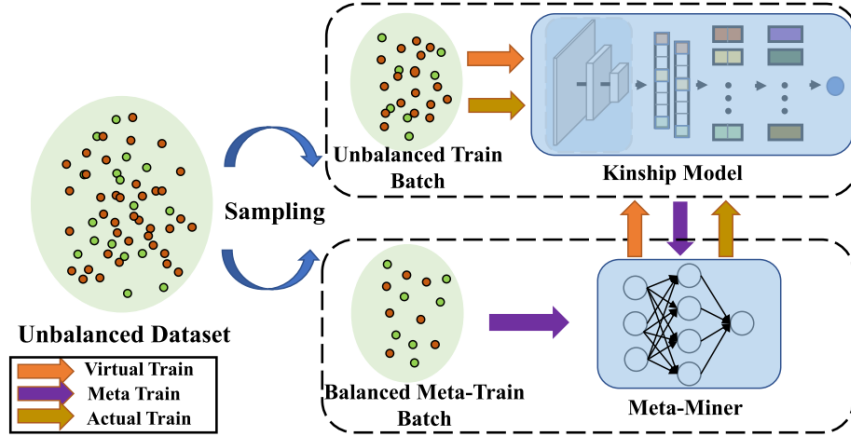


Figure 14: Framework of a meta-learning based network, cited from Li et al. [113].

message passing scheme is employed. The final results are computed by a predefined MLP.

**Meta-learning:** Deep learning-based methods show good performance in solving extrinsic challenges. One of the extrinsic challenges lies in that "Kinship verification databases are born with unbalanced data" [113]. A kinship dataset of  $N$  pairs of positive samples contains  $N(N - 1)$  potential negative pairs leading to a large unbalance. However, most of the current methods only use  $N$  negative pairs. Recently, Li et al. [113] propose a Discriminative Sample Meta-Mining (DSMM) approach to exploit all possible pairs and learn discriminative information. As depicted in Figure 14, a meta-miner is deployed to mine the distinctive samples by re-weighting the sample ratios in the training batch with a meta-gradient. This framework simultaneously samples two training batches with different ratios. It conducts sample mining on the training batch under the guidance of the balanced meta-train batch.

#### Others

There are two types of transfer learning: inductive and transductive transfer learning. For inductive transfer learning, the distribution of learning targets can be different [73, 233]. In contrast, transductive transfer learning always keeps the learning target identical, but the embedded data distribution is often changed. Two transductive transfer learning

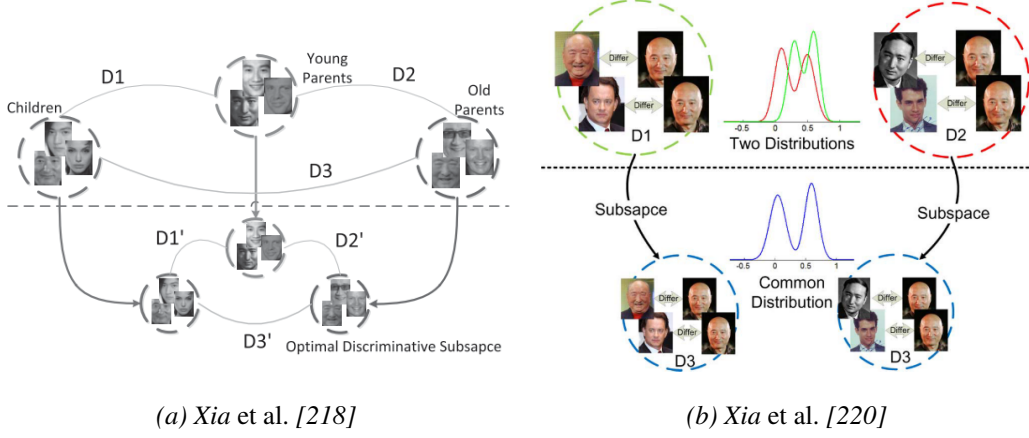


Figure 15: Transfer subspace learning methods [218, 220] for kinship verification.  $D1$ ,  $D2$ , and  $D3$  correspond to target domain, source domain, and learned subspaces, cited from Xia et al. [220].

methods are proposed by Xia *et al.* [218, 220] aiming to improve the representation of latent features.

**Transductive transfer learning:** Xia *et al.* [218] use intermediate data by collecting images of young parents as intermediate sets. Both the source and target distributions are supposed to be close to the intermediate set to yield a similar distribution and are formulated as follows:

$$W = \arg \min_W \{F(W) + \lambda_1 D_W(P_L \| P_U) + \lambda_2 D_W(P_L \| P_V)\}, \quad (9)$$

where the  $F(W)$  is a general subspace learning (*e.g.*, PCA [188], LDA [12] and DLA [243]). The distribution of the source, intermediate, and target set correspond to  $P_U$ ,  $P_L$  and  $P_V$ , respectively.  $D_W(P_L \| P_U)$  and  $D_W(P_L \| P_V)$  are Bregman divergence-based regularization. In fact, the intermediate set becomes the bridge to connect the other two sets. Xia *et al.* simplify the method by using two distributions based on pairwise differences instead of transferring three distributions together to a general subspace, as defined by:

$$W = \arg \min_W \{F(W) + \lambda D_W(P_L \| P_U)\}. \quad (10)$$

As shown in Figure 15, the task corresponds to finding a subspace, where the different distribution of the two pairs (child-young parent and child-old parent) has a similar distribution while keeping distinctions.

**Inductive transfer learning:** Inductive transfer learning is often used in deep learning methods, exploiting the feature-extracting capability of the pre-trained neural net model. Robinson *et al.* [166] use several methods and benchmark them on the FIW data. The pre-trained convolutional neural network is taken as an on-the-shelf feature extractor. Specifically, the layers of the pre-trained VGG-Face model are frozen, except for the second-to-last fully-connected layer.

### 2.4.2 Video-based

By the end of 2017, existing kinship verification methods are mainly based on static images. However, important kinship-related information can be derived from facial dynamics/motion. For example, children may have similar facial expressions as their parents such as smile, anger, astonishment, *etc.*) [49]. Research [152] also shows that parents and children have genetic similarities in facial dynamics. Obviously, static images do not provide such information *i.e.*, pose variations, facial expression changes, dynamic movement, adequate 3D estimation, *etc.* Hence, video-based kinship datasets are required.

#### *Handcrafted descriptors*

Dibeklioglu *et al.* [49] is the first to use a video dataset for kinship verification. They exploit dynamic information from smiling using the UvA-NEMO Smile dataset. First, the displacement of eyebrows, eyelids, cheeks, and lip corners are computed based on the movement of landmark points. Then, spatio-temporal features are extracted using the temporal Completed Local Binary Pattern (CLBP) descriptors from multiple frames. Finally, the combined information from dynamic and spatio-temporal features are used by a SVMs jointly. After the publication, several other methods are proposed. Boutellaa *et al.* [18] use the shallow spatio-temporal texture information and deep information. Yan *et al.* [226] collect a new dataset (KFVW) and evaluate different metric learning methods.

#### *Metric learning*

Yan *et al.* [226] evaluate a number of metric learning-based methods using the KFVW dataset. One hundred frames are randomly extracted from each video with a cropped face region. Then, all images are converted to gray-scale. LBP features and HOG features are extracted for comparison. Information-theoretic metric learning (ITML), side-information-based linear discriminant analysis (SILD), KISS metric learning (KISSME), and cosine similarity metric learning (CSML) are evaluated. The final results show that the LBP feature obtains better performance than using HOG.

#### *Deep learning*

**DeepFeat:** Inspired by [49], Boutellaa *et al.* [18] use spatio-temporal information for video-based kinship verification. Instead of using handcrafted features, they use pre-trained VGG-Face in an off-the-shelf way to extract features. The spatio-temporal features are extracted by three different handcrafted methods: LBPTOP, LPQTOP, and BSIFTOP. The results show that combining shallow with deep features obtains the best results.

**SMNAE:** [100] Kohli *et al.* [101] propose a deep learning framework for kinship verification in unconstrained videos using Supervised Mixed Norm Autoencoders (SMNAE). This auto-encoder formulation introduces class-specific sparsity in the weight matrix.

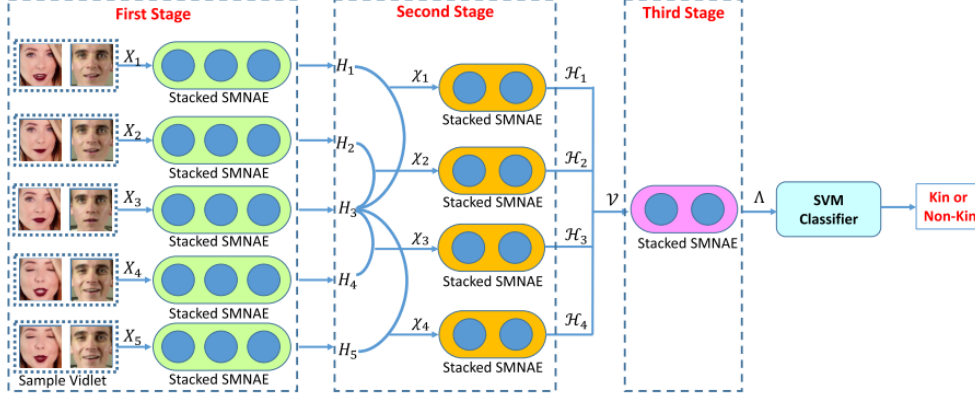


Figure 16: The three-stage kinship verification in unconstrained videos framework by using SMNAE, cited from Kohli et al. [100].

The Mixed Norm Auto-encoder (SMNAE) combines  $l_{2,p}$  norm and a pairwise class-based sparsity penalty with loss function  $J_{SMNAE}$  formulated by:

$$J_{SMNAE} = \arg \min_{\mathbf{W}, \mathbf{W}'} \left\| \mathbf{x} - \phi(\mathbf{W}' \mathbf{H}) \right\|_F^2 + \lambda \sum_{c=1}^C \left\| \mathbf{W} \mathbf{x}_c \right\|_{2,p} + \beta \left( \text{tr}(\mathbf{H}^T \mathbf{H} \mathbf{L}) \right), \quad (11)$$

where  $\mathbf{W}$  is the weight matrix and  $\phi$  is the activation function.  $\mathbf{L}$  is the Laplacian matrix, which can be taken as  $\mathbf{L} = \mathbf{D} - \mathbf{M}$ .  $\mathbf{D}$  is the diagonal matrix, and  $\mathbf{M}$  is the adjacency matrix. They use this formulation to develop a three-stage framework. The framework of SMNAE is illustrated in Figure 16. In the first stage, the video pair is split into non-overlapping vidlets. These vidlets are fed into a stacked SMNAE to yield a spatial representation. In the second stage, the learned spatial representations are concatenated pair-wisely and then fed into the second stage's stacked SMNAE. The third stage mainly receives the global Spatio-temporal information. The encoded representation is used by an SVM for the final classification. The aim of the approach is to obtain spatial and temporal information by using an auto-encoder, resulting in a discriminative but sparse representation.

### 2.4.3 Multi-label Methods

**Audio:** As mentioned in [216], the University of Nottingham's experiment shows that the human voice contains heritable information. Other research also shows that the voice of people with close kin relationship results in a performance degradation of automatic speaker verification (ASV) [10, 105]. Inspired by this observation, assuming that the human voice contains kin-related cues, Wu *et al.* [216] fuse face and voice modalities to improve the accuracy and robustness of kinship systems verification. They propose a Siamese fusion network with a contrastive loss utilizing the fine-tuned VGG-Face CNN cascaded and an LSTM network. To extract voice features, they pre-train a ResNet-50 [150] on VoxCeleb2 [35] and fine-tune it with TALKIN. These two models are trained

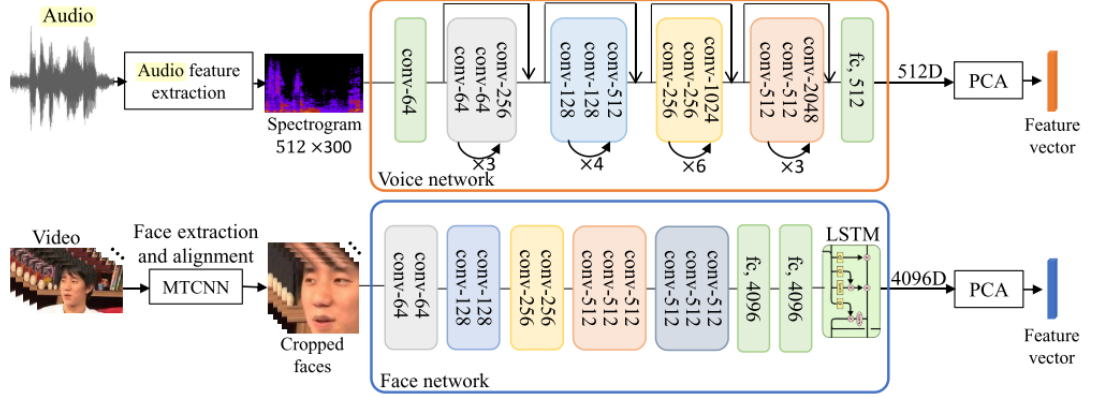


Figure 17: Architecture of an audio-based method, cited from Wu et al. [150].

using a contrastive loss to learn intra-class similarity and inter-class dissimilarity. After feature extraction, PCA is used to reduce the feature dimensions for both face and audio. Facial and vocal features are reduced to 130 and concatenated together to form a 260-dimensional feature. After the  $fc$  layer, the outcome is evaluated by using the cosine similarity. The results show that the vocal information improves the accuracy by around 3 percent.

**Age:** Previous studies [8, 38, 93, 218, 220] show that changes in age may negatively affect the accuracy of kinship verification. Because of the age gap, the parents' face structures are deformed compared to the face when they were young [161]. It indicates the possibility of improving the accuracy of transforming people's facial information by age. Similar ideas are proposed by Xia [218, 220] to transfer the distribution of children and parents to a general subspace, which indirectly utilizes the age information. In addition, Wang *et al.* [197] generate parent images to their younger ages, and Dehshibi [45] propose an age-aware facial kinship verification to fill the gap of aging effects in asserting kin-relation.

**Graph based:** Xia *et al.* [220] assume that people have a higher kin likelihood when they are located together in an image. For example, in a family photo, senior people often sit in the middle surrounded by their family members. The paper utilizes the information to improve kinship verification by combining relative distance, gender relation, age difference, and kinship score. In [78], the potential relationships of people in one photo are transformed into a set of candidate graphs with all possible relationships. Then, they accumulate the scores of each candidate graph. The graph with the highest score corresponds to the final kinship prediction.

#### 2.4.4 Discussion

Details of the different methods are listed in Table 2 and Table 3. Most methods focus on solving the 4-types of kinship verification task using public datasets collected online in an unconstrained environment. Only a few methods focus on kinship types with two generation skips (11 types). The Family 101, KinFaceW-I&II, and CornellKin and the metric-learning-based methods are mostly used. Many of the metric-learning-



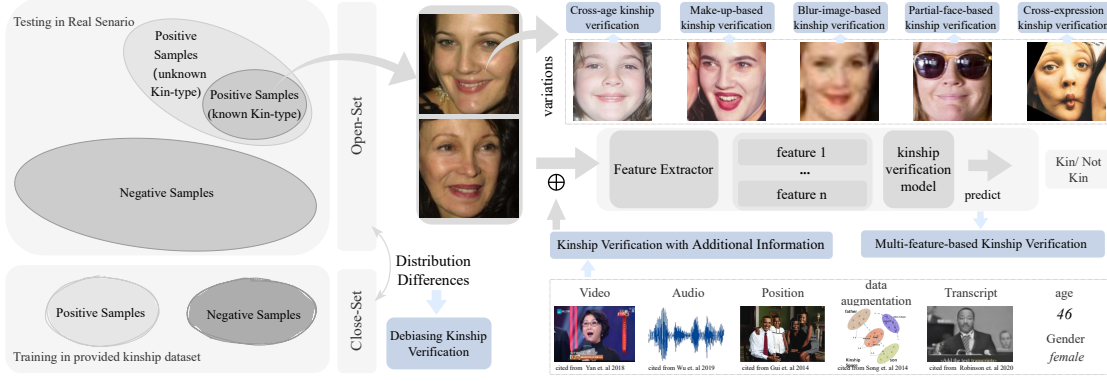


Figure 18: Potential directions for kinship verification.

based methods, obtaining high accuracy, follow a similar strategy: they have multiple descriptors with different ranges of scales and deeper descriptors. On the other hand, some methods focus on specific challenges. As shown in Figure 8, the challenge of unconstrained-images-based kinship verification is often approached by data-driven methods. However, only a few methods focus on solving the unconstrained challenges based on a specific design. For example, to our knowledge, there is no approach to adjust the methods to deal with pose variations and occlusion problems. As for intrinsic challenges, expression changes are mostly taken into account for video-based kinship verification. Recently, many methods focus on utilizing gender information, whereas ethnicity’s side-effect is largely ignored so far. Also, for the age differences, there are several methods that focus on solving larger age differences and old-parents-related tasks. However, a few methods focus on how to deal with children-related pairs. The graph of methods shows that there are still many unsolved problems. From the milestones in Figure 9, it is shown that more and more deep learning methods are used. There is also a trend to combine metric learning with deep learning.

#### 2.4.5 Potential Directions for Kinship Verification

Kinship verification is a challenging but promising task. Currently, there are still many open directions, see Figure 18. For example, most of the current kinship verification methods are close-set approaches. Both testing and training data are from the same kin-type set. However, this evaluation protocol ignores many unknown relationships in real-world scenarios and omits the influence of other kin-type samples. Conducting kinship verification in an open-set environment is a promising direction. Another point is that there is a racial imbalance in the data collection and construction process. Positive and negative samples also do not match real-world scenarios. Debiasing kinship verification is of great importance. In addition, there are many interference factors. Research in cross-age kinship verification, cross-expression kinship verification, make-up-based kinship verification, and partial-face-based kinship verification is required. Due to the development of deepfakes [209], anti-spoofing becomes important. Increasing the kinship system’s stability is a promising direction. How to combine various types of data and features at different levels is still an unsolved problem.

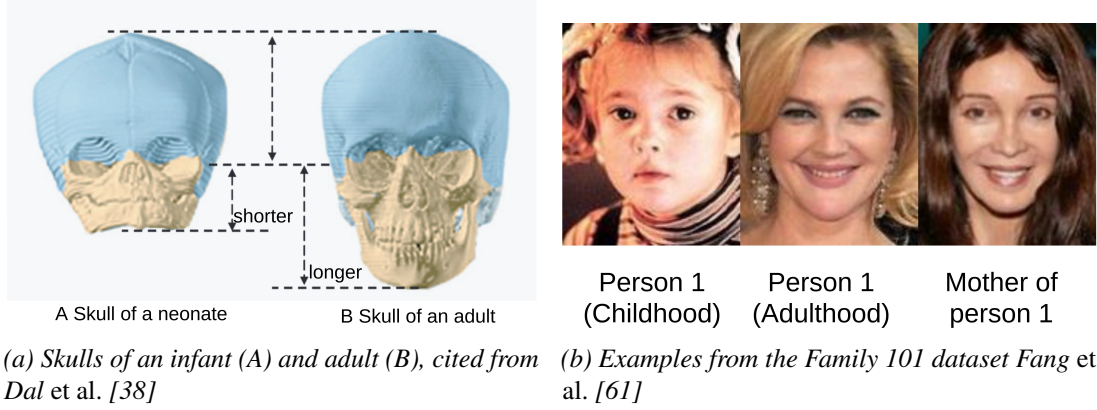


Figure 19: (a) Differences in face outlines between children and adults increase the appearance discrepancy. (b) Variations in skin color and texture between childhood and adulthood increase the heterogeneity of the same person.

## 2.5 NEMO-DATASET

### 2.5.1 Motivation

Kinship verification based on child-adult-related pairs is useful as child-adult pairs often occur in many applications such as children’s adoption and missing children searching. The performance of kinship verification on child-adult pairs is negatively influenced by large variations among children and adults. As shown in Figure 19, the facial outlines of the same person during childhood and adulthood change drastically.

However, only a few researchers focus on child-adult kinship verification. One reason is the shortage of child-adult images-based kinship datasets. Considering the commonality of public datasets, this specific task for kinship verification cannot be studied systematically.

### 2.5.2 Data Collection

The aim of the Nemo-Kinship Dataset is to collect child-adult-based kinship-related videos with multiple labels. To this end, we collect the kinship-related data from a deception-testing experiment at Nemo Museum as part of the scientific experiments of NEMO Science Live Program<sup>8</sup>.

#### *Recording conditions*

During the data collection process, the participants are divided into different groups based on family or friend relationships, and the language they speak. Participants in each group take turns to undergo the experiment as test subjects. According to the allocated questions, all the participants’ answers during the experiments are recorded and divided into 13 different video clips. The entire experiment is recorded by a web

<sup>8</sup> <https://www.nemosciencemuseum.nl/nl/wat-is-er-te-doen/activiteiten/science-live/>

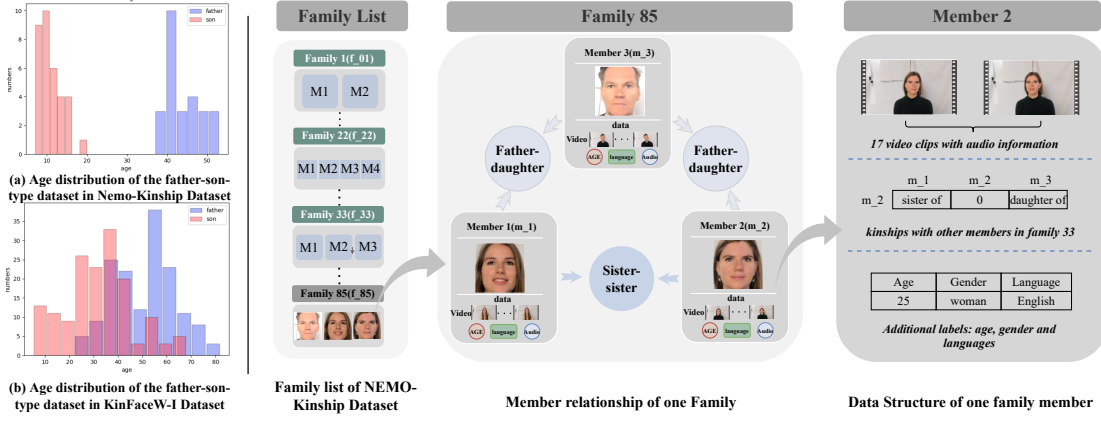


Figure 20: Structure of the proposed Nemo-Kinship dataset and age distribution. On the left side: (a): Age distribution of the Nemo-kinship dataset. (b): Age distribution of the KinFaceW-I dataset is annotated by two participants manually, since there is no age label in KinFaceW-I dataset. On the right side: The Nemo-Kinship images are stored in a family list.

camera connected to a computer. The web camera records video information together with audio. The video has a resolution of  $1920 \times 1080$  pixels at a speed of 60 frames per second, and the audio codec format is *MPEG – 4AAC* with *stereo* channel,  $48000\text{Hz}$  of sample rate and  $320\text{kbps}$  of Bit-rate. During the entire experiment, the camera’s angle and the position are kept the same, so all test subjects are taken with a frontal view in a controlled environment. Incandescent lamps are arranged around the entire interview room to ensure that the light is as stable as possible to eliminate the interference caused by environmental changes.

### Data annotation

After collecting all the practical information of participants, 248 participants with kinship-related information are kept. An auto-clipping tool is created to divide the video into separate clips according to the interview questions and answers. Each video is kept along with audio information containing the speech. There are 11 valid kinship types: F-D, F-S, M-D, M-S, B-B, B-S, S-S, GF-GD, GF-GS, GM-GD, GM-GS. Also, age, gender, and family relationship information is collected. As depicted in Figure 20, the videos of the participants are arranged in separate family units ( $f_{.01} - f_{.85}$ ). All family members are stored sequentially ( $m_1, m_2, \dots$ ). Each family member contains 17 video clips. Each family member’s label contains age, gender, spoken language, and kinship relationships between other members in the same family unit.

### 2.5.3 Data Statistics

The dataset consists of a large proportion of children’s videos, making it easier to focus on child-adult-related kingship verification tasks. It contains 4216 videos of 248 family members from 85 families. It contains 11 kinship types and the age of each family member. The ages of the family members vary from 7 to 71. A child is a person whose age is under 16. The age distribution of the dataset compared to KinFaceW-I is depicted

Table 4: Statistics of Nemo-Kinship dataset.

Kin-type	parent-child				siblings			grandparent-grandchild			
	F-S	F-D	M-S	M-D	S-S	B-B	B-S	GM-GS	GM-GD	GF-GD	GF-GS
pairs	34	30	42	46	15	15	31	5	6	3	3
Children related	33	26	40	44	12	14	28	5	6	3	3
Family numbers	26	25	34	37	15	13	26	2	5	3	2
English speaker	20	15	18	20	10	6	8	0	0	2	0
Dutch speaker	40	40	57	64	20	21	49	6	10	4	5
Male	60	25	40	0	0	27	28	4	0	3	5
Female	0	30	35	84	30	0	29	2	10	3	0
individuals	60	55	75	84	30	27	57	6	10	6	5
Total individuals	248										

in Figure 20. Statistics of the Nemo-Kinship dataset are shown in Table 4. A comparison between the Nemo-Kinship dataset and other related datasets is listed in Table 5.

In conclusion, the Nemo-Kinship dataset:

1. contains a large proportion of children images.
2. consists of different labels and can be used for different tasks such as:
  - Video-based kinship verification.
  - Child-adult-based kinship verification.
  - Kinship verification with age information.
  - Kinship verification with gender information.
  - Kinship verification with audio information.
  - Family classification.
  - Kin-based detection.

## 2.6 EVALUATION PROTOCOLS, METRICS FOR KINSHIP VERIFICATION

### 2.6.1 Protocols

The kinship verification task is a binary classification problem. The mostly used evaluation protocol for kinship verification is K-fold cross-validation. Because the number of samples of the training dataset is limited, the test results may vary. K-fold cross-validation provides relatively more stable results. The mostly used cross-validation is 5-fold cross-validation. Using the same cross-validation fold provides a comparison between the proposed and methods in the literature.

### 2.6.2 Metrics

The mostly used evaluation metrics are classification accuracy [70] and EER. Each kinship's result is calculated and obtained for classification accuracy by dividing the correct number by the total test number. Finally, an accuracy metric is obtained with

Table 5: Details of the different kinship datasets. Video<sup>†</sup> denotes that the videos are recorded with spontaneous smiles. Sizes with \* indicate that the original size is  $1920 \times 1080$ . Brackets indicate the updated number compared to the original numbers collected.

Year	Name	types	pairs/ groups	kin types	subjects	numbers	size	extra labels	family structure	age	time range	source	top method	score	Human score
2010	CornellKin [62]	images	150 (143)	4	300*	300*	100x100	no	no	no	no	online	Kohli <i>et al.</i> [101]	94.4	67.19
2012	UB KinFacev1 [218]	images	90 *	-	180	270	-	young, old parents	no	no	yes	online	Yan <i>et al.</i> [230]	67.3	-
2012	UB KinFacev2 [176]	images	200*	4	400	600	127x100	young, old parents	no	no	yes	online	Kohli <i>et al.</i> [101]	95.3	54.85
2012	familyFace [221]	images	-	4	507	214	-	family tree, age, position	yes	no	yes	online	-	-	-
2013	Family101 [61]	images	692 (206)	4	607	14,816	-	family tree, 206 nuclear families	yes	no	yes	Amazon MTurk	wang <i>et al.</i> [205]	92.03	-
2013	UvA-Nemo Smile [49]	videos <sup>†</sup> videos (posed)	228 287	7 7	75 87	456 564	125x100*	video	yes	no	no	offline	Kohli <i>et al.</i> [101]	96.07	80.2
2014	KinFaceW-I [129]	images	533	4	1066	1066	64x64	-	no	no	no	online	Kohli <i>et al.</i> [101]	96.9	78.6
2014	KinFaceW-II [129]	images	1000	4	2000	2000	64x64	-	no	no	no	online	Kohli <i>et al.</i> [101]	97.1	83.5
2015	TSKinFace [158]	images	1015	3 (2)	2589	787	-	family tree	yes	no	yes	online	zhang <i>et al.</i> [239]	89.8	79.55
2016	FIW [168]	images	656, (954)	11	10,676	30,725	224x224	family tree	yes	no	yes	online	Robinson <i>et al.</i> [168]	69.18	57.5
2017	KFWV [226]	videos	418	4	-	-	900x500	video	no	no	no	online TV show	yan <i>et al.</i> [226]	58.15	73
2019	KIVI [101]	videos	355	7	503	503*	128x128	video	yes	no	no	online	Kohli <i>et al.</i> [101]	83.18	-
2019	TALKIN [216]	videos	400	4	-	-	224x224*	video,audio	no	no	no	online TV show	wu <i>et al.</i> [216]	70.2	-
2019	Nemo- Kinship (ours)	videos	228	11	248	4216	160x160*	family tree, age, language audio	yes	yes	no	offline	-	-	-

an average score for each subset. Besides classification accuracy, receiver operating characteristic (ROC) curves are used.

## 2.7 BENCHMARKING

Table 6: Results of methods on 4-type relation datasets. The results with <sup>†</sup> and <sup>‡</sup> indicate the kinship verification performance on the children young-parents related set and children old-parents related set respectively for UB KinFace.

Dataset Format	Dataset	Method	Description	F-S	F-D	M-S	M-D	average
CornellKin		Fang <i>et al.</i> 2010 [62]	kNN,SVM	-	-	-	-	70.7
		Yan <i>et al.</i> 2014 [228]	DMML,SVM	76.0	70.6	77.5	71.0	-
		Lu <i>et al.</i> 2013 [130]	MNRML,SVM	74.5	68.8	77.2	65.8	71.6
		Yan <i>et al.</i> 2015 [230]	MPDFL,SVM	74.8	69.1	77.5	66.1	71.9
		Liu <i>et al.</i> 2016 [121]	GInCS	78.2	73.0	78.8	73.5	75.9
		kohli <i>et al.</i> 2017 [99]	fcDBN,3-1NN	91.7	87.9	95.2	84.2	-
		Dehshibi <i>et al.</i> 2019 [45]	K-BDPCA	74.8	69.1	77.5	66.1	-
		Zhou <i>et al.</i> 2019 [250]	kinship metric learning	78.9	82.6	78.3	85.7	81.4
		Serraoui <i>et al.</i> 2022 [174]	TXQEDA+WCCN	-	-	-	-	93.77
UB KinFace		Shao <i>et al.</i> 2011 [176]	kNN,CMC	-	-	-	-	56.5
		Xia <i>et al.</i> 2011 [219]	Region(Gabor),DLA,Transfer Learning	-	-	-	-	60.0
		Xia <i>et al.</i> 2012 [221]	Local Gabor, TSL	-	-	-	-	56.5
		Yan <i>et al.</i> 2014 [228]	DMML,SVM	74.5 <sup>†</sup>	-	70.0 <sup>‡</sup>	-	-
		Lu <i>et al.</i> 2014 [130]	MNRML,SVM	67.3 <sup>†</sup>	-	66.8 <sup>‡</sup>	-	-
		Yan <i>et al.</i> 2015 [230]	MPDFL,SVM	67.5 <sup>†</sup>	-	67.0 <sup>‡</sup>	-	67.3
		Liu <i>et al.</i> 2016 [121]	GInCS	75.8 <sup>†</sup>	-	72.2 <sup>‡</sup>	-	-
		Kohli <i>et al.</i> 2017 [99]	fcDBN	92.0 <sup>†</sup>	-	91.5 <sup>‡</sup>	-	-
		Dehshibi <i>et al.</i> 2019 [45]	K-BDPCA	72.5	66.5	66.2	72.0	-
		Zhou <i>et al.</i> 2019 [250]	kinship metric learning	75.8 <sup>†</sup>	-	75.2 <sup>‡</sup>	-	75.5
	Family101	Fang <i>et al.</i> 2013 [61]	sparse group lasso	-	-	-	-	32
		Wang <i>et al.</i> 2014 [205]	GMM,SVM	92.3	-	-	-	-
		Dehshibi <i>et al.</i> 2019 [45]	K-BDPCA	86.8	82.8	84.4	83.2	-
		Lu <i>et al.</i> 2014 [130]	MNRML,SVM	72.5	66.5	66.2	72.0	69.9
		Yan <i>et al.</i> 2014 [228]	DMML,SVM	74.5	69.5	69.5	75.5	-
		Dehghan <i>et al.</i> 2014 [44]	gated autoencoder	76.4	72.5	71.9	77.3	74.5
		Liu <i>et al.</i> 2015 [120]	IFVF	73.4	71.7	71.1	77.6	73.5
		Alirezazadeh <i>et al.</i> 2015 [5]	genetic algorithm	77.9	78.0	81.4	87.9	81.3
		Zhang <i>et al.</i> 2015 [240]	CNN-point	76.1	71.8	78.0	84.1	77.5
		Bottinok <i>et al.</i> 2015 [17]	handcrafted features, svm	85.8	85.3	86.7	86.7	86.3
		Duan and Tan 2015 [57]	LPQ	75.4	63.8	69.9	74.6	70.9
		Yan <i>et al.</i> 2015 [230]	MPDFL,SVM	73.5	67.5	66.1	73.1	70.1
		Xu <i>et al.</i> 2016 [223]	S3L	82.4	72.8	74.6	79.1	77.2
		Zhou <i>et al.</i> 2016 [253]	Multiview SSL	82.8	75.4	72.6	81.3	78.0

	KinFaceW-I	Zhou <i>et al.</i> 2016 [252]	Ensemble similarity learning	83.9	76.0	73.5	81.5	78.6
		Puttenputhussery <i>et al.</i> 2016 [155]	SF-GFVF feature	76.3	74.6	75.5	80.8	76.1
		Li <i>et al.</i> 2016	SMCNN	75.0	75.0	72.2	68.7	-
		Liu <i>et al.</i> 2016 [121]	GInCS	77.3	76.9	75.8	81.4	77.8
		Liang <i>et al.</i> 2017 [116]	autoencoder,SVM	71.2	74.3	77.2	73.3	-
		Lu <i>et al.</i> 2017 [128]	DDMML	86.4	79.1	81.4	87.0	83.5
		Yan 2017 [224]	NRCML	73.4	70.6	70.8	69.9	73.1
		Patel <i>et al.</i> 2017 [151]	BNRML	83.38	77.25	75.80	78.36	78.7
		Fang <i>et al.</i> 2017 [63]	SSML	84.6	75.0	76.3	82.3	79.6
		kohli <i>et al.</i> 2017 [99]	fcDBN	98.1	96.3	90.5	98.4	-
		Dehshibi <i>et al.</i> 2019 [45]	K-BDPCA	77.9	78.0	81.4	87.9	-
		Dibeklioglu <i>et al.</i> 2017 [48]	deep contrastive learning	-	-	-	-	80.5
		Chen <i>et al.</i> 2017 [29]	multi-linear coherent space learning	88.5	81.0	81.0	82.6	83.3
		Mahpod and Keller 2018 [134]	MHDL3	77.0	76.1	80.1	85.8	79.8
		Kohli <i>et al.</i> 2018 [100]	Supervised Mixed Norm Autoencoder	-	-	-	-	96.9
		Liang <i>et al.</i> 2019 [117]	WGEML	78.5	73.9	80.6	81.9	78.7
		Zhou <i>et al.</i> 2019 [250]	kinship metric learning	83.8	81.0	81.2	85.0	82.8
		Yan 2019 [225]	D-CBFD	79.6	73.6	76.1	81.5	77.6
		Li <i>et al.</i> 2020 [114]	graph-based kinship reasoning	79.5	73.2	78.0	86.2	79.2
		Serraoui <i>et al.</i> 2022 [174]	TXQEDA+WCCN	91.00	87.78	92.32	93.35	91.11
Image	KinFaceW-II	Lu <i>et al.</i> 2014	MNRLM,SVM	76.9	74.3	77.4	77.6	76.5
		Yan <i>et al.</i> 2014 [229]	DMML,SVM	78.5	76.5	78.5	79.5	78.3
		Dehghan <i>et al.</i> 2014 [44]	gated autoencoder	83.9	76.7	83.4	84.8	82.2
		Hu <i>et al.</i> 2014 [86]	LM3L	82.4	74.2	79.6	78.7	78.7
		Liu <i>et al.</i> 2015 [120]	IFVF	85.6	75.4	82.8	82.6	81.6
		Alirezazadeh <i>et al.</i> 2015 [4]	genetic algorithm	88.8	81.8	86.8	87.2	86.2
		Zhang <i>et al.</i> 2015 [240]	CNN-point	89.4	81.9	89.9	92.4	88.4
		Bottinok <i>et al.</i> 2015 [17]	handcrafted features, SVM	89.4	83.6	86.2	85.0	86.0
		Duan and Tan 2015 [57]	LPQ	82.4	76.2	76.6	73.2	77.1
		Yan <i>et al.</i> 2015 [230]	MPDFL,SVM	77.3	74.7	77.8	78.0	77.0
		Xu <i>et al.</i> 2016 [223]	S3L	82.6	73.8	74.1	73.6	76.0
		Zhou <i>et al.</i> 2016 [253]	Multiview SSL	81.8	74.0	75.3	72.5	75.9
		Zhou <i>et al.</i> 2016 [252]	Ensemble similarity learning	81.2	73.0	75.6	73.0	75.7
		Puttenputhussery <i>et al.</i> 2016 [155]	SF-GFVF feature	87.2	79.6	88.0	87.8	85.7
		Li <i>et al.</i> 2016 [111]	SMCNN	79.0	75.0	85.0	78.0	-
		Liu <i>et al.</i> 2016 [121]	GInCS	85.4	77.0	81.6	81.6	81.4
		Liang <i>et al.</i> 2017 [116]	autoencoder,SVM	80.3	85.2	83.3	82.6	-
		Lu <i>et al.</i> 2017 [128]	DDMML	87.4	83.8	83.2	83.0	84.3
		Yan 2017 [224]	NRCML	79.8	76.1	79.8	80.0	78.7
		Patel <i>et al.</i> 2017 [151]	BNRML	84.0	79.0	79.2	80.0	80.6
		Hu <i>et al.</i> 2017 [85]	L2M3L	82.4	78.2	78.8	80.4	80.0
		Fang <i>et al.</i> 2017 [63]	SSML	85.0	77.0	80.4	78.4	80.2
		kohli <i>et al.</i> 2017 [99]	fcDBN	96.8	94.0	97.2	96.8	-
		Hu <i>et al.</i> 2018 [84]	MvDML	80.4	79.8	78.8	81.8	80.2
		Dehshibi <i>et al.</i> 2019 [45]	K-BDPCA	88.8	81.8	86.8	87.2	-
		Dibeklioglu <i>et al.</i> 2017 [48]	deep contrastive learning	-	-	-	-	82.3
		Chen <i>et al.</i> 2017 [29]	multi-linear coherent space learning	86.8	82.8	84.4	83.2	84.3
		Mahpod and Keller 2018 [134]	MHDL3	88.4	84.0	86.4	89.2	87.0
		Kohli <i>et al.</i> 2018 [100]	Supervised Mixed Norm Autoencode	-	-	-	-	97.1
		Liang <i>et al.</i> 2019 [117]	WGEML	88.6	77.4	83.4	81.6	82.8
		Zhou <i>et al.</i> 2019 [250]	kinship metric learning	87.4	83.6	86.2	85.6	85.7
		Yan 2019 [225]	D-CBFD	81.0	76.2	77.4	79.3	78.5
		Li <i>et al.</i> 2020 [114]	graph-based kinship reasoning	90.8	86.0	91.2	94.4	90.6
		Serraoui <i>et al.</i> 2022 [174]	TXQEDA+WCCN	89.80	90.60	87.60	93.20	90.30
	TSKinFace	Qin <i>et al.</i> 2015 [158]	RSBM	83.0	80.5	82.8	81.1	FM-S: 86.4, FM-D: 84.4
		Zhang <i>et al.</i> 2016 [239]	HDLBPH	91.1		88.3		89.7
		Lu <i>et al.</i> 2017 [128]	DDMML	86.6	82.5	83.2	84.3	FM-S: 88.5, FM-D: 87.1
		Zhang <i>et al.</i> 2015 [244]	GMP	88.5	87.0	87.9	87.8	FM-S: 90.6, FM-D: 89.0
		Liang <i>et al.</i> 2019 [117]	WGEML	90.3	89.8	91.4	90.4	FM-S: 93.5, FM-D: 93.0
		Serraoui <i>et al.</i> 2022 [174]	TXQEDA+WCCN	89.42	89.31	90.87	93.15	FM-S: 95.34, FM-D: 96.53
Video	KFVW	Yan <i>et al.</i> 2018 [226]	CSML	38.6*	47.1*	38.5*	43.2*	41.8*
		Yan 2019 [225]	D-CBFD	61.5	57.0	58.8	59.9	59.3
		Wu <i>et al.</i> 2019 [216]	Deep Siamese Network	80.0	70.5	73.5	72.5	74.1

### 2.7.1 Benchmark on Publicly Available Datasets

#### Relation degree

**4 types:** Most of the current methods train and test on 4-types datasets. The performances of the different methods based on the different 4-types datasets are listed in Table 6. It can be concluded that the video dataset is more complicated than the image dataset, since the highest accuracy of the KFVW dataset is 59.3%, and for TALKIN it is 74.1%. UB KinFace yields a lower performance considering image-based datasets when testing this 4-type using the same method. MNRLM achieves an accuracy of 67.1 for UB KinFace and 71.6, 69.9, and 76.5 for CornellKin, KinFaceW-I, and KinFaceW-II, respectively. KinFaceW-I and KinFaceW-II are the mostly used public datasets. The traditional metric learning methods MNRLM and DMML show relatively low performance. Compared to MNRLM, DDMML performs better by utilizing multiple neural networks and using the commonality of multiple feature descriptors. Based on the comparison, deeper features usually result in higher accuracy. For instance, fcDBN uses a convolutional neural network and SMNAE uses mixed norm auto-encoders. it can be concluded that multi-descriptors provide better results. Compared to NRML and DDML, which only use one specific feature descriptor, MNRML and DDMML achieve better results on the KinFaceW-I&II datasets. A partial-aware feature extractor is also helpful. The Attention Network uses a part-aware attention module. fcDBN uses hierarchical representations with local and global facial regions. Both of them show good performance. In conclusion, deeper feature extractors, multi-descriptors, and part-aware extractors are useful for kinship verification.

Table 7: Results of Methods on 7-type relation Datasets.

Dataset Format	Dataset	Method	Description	F-D	F-S	M-D	M-S	B-B	B-S	S-S	average
Image	WVU Kinship Db IIITD-Kinship	Kohli <i>et al.</i> 2017 [99]	fcDBN	88.4	90.8	95.2	90.6	90.9	87.5	95.7	-
		Kohli <i>et al.</i> 2012 [98]	SSD, SVM	73.4	77.3	79.6	71.0	73.0	68.7	78.3	-
		Dibeklioglu <i>et al.</i> 2013 [49]	dynamic features, CLBP-TOP	75.0	79.0	67.54	75.0	70.0	68.8	75.0	72.9
Video	UvA-NEMO Smile	Boutellaa <i>et al.</i> 2016 [18]	DeepFeat	89.7	92.7	90.2	85.7	92.8	88.5	88.9	89.8
		Dibeklioglu 2017 [48]	contrastive learning	93.8	93.4	93.6	92.2	95.7	92.6	94.2	93.6
		Kohli <i>et al.</i> 2019 [100]	SMNAE	-	-	-	-	-	-	-	83.2

**7 types:** Table 7 shows the performances of different methods on the 7-type kinship datasets. Compared to 4-type datasets, fewer methods focus on 7-type datasets. To our knowledge, fcDBN is the only method tested on the WVU kinship dataset. As for the video-based dataset, the UvA-NEMO Smile dataset is the most widely used. The best performance on this dataset is 93.6%. Among these methods, Dibeklioglu *et al.* [49] reach an average accuracy of 72.9. Boutellaa *et al.* [18], achieve an average accuracy of 89.8. Dibeklioglu *et al.* [49] use traditional descriptors by combining facial dynamics and spatio-temporal appearance. Boutellaa *et al.* [18] use the spatio-temporal information and use deep features from VGG-face.

**11 types:** FIW is the largest image-based dataset *al.l* methods, conducted on the FIW dataset, are based on neural network architectures.

Table 8: Results of the methods on 11-type relation datasets.

Dataset Format	Dataset	Method	Description	F-D	F-S	M-D	M-S	B-S	B-B	S-S	GF-GD	GF-GS	GM-GD	GM-GS	average
Image	FIW	Robinson <i>et al.</i> 2016 [167]	VGG-face+SVM	64.4	63.4	66.2	64.0	73.2	71.5	70.8	64.4	68.6	66.2	63.5	66.9
		Robinson <i>et al.</i> 2016 [166]	Fine-Tuned CNN	69.4	68.2	68.4	69.4	74.4	73.0	72.5	72.9	72.3	72.4	68.3	71.0
		Wang <i>et al.</i> [198]	VGG+DML	68.1	71.0	70.4	70.8	75.3	-	-	64.9	64.81	67.4	66.5	68.8
	FIW (extended)	Robinson <i>et al.</i> 2017 [170]	ResNet+Centerface	68.2	67.7	71.1	68.6	69.5	69.9	69.5	66.4	66.5	65.8	64.4	67.9
		Robinson <i>et al.</i> 2018 [168]	SphereFace	69.3	68.5	71.8	69.5	70.2	71.9	77.3	66.1	66.36	64.6	65.4	69.2
		Laiadi <i>et al.</i> 2019 [106]	Deep-Tensor+ELM	-	-	-	-	-	-	-	68.4	68.2	70.2	67.8	68.6
		Wang <i>et al.</i> 2018 [197]	ResNet+SDMLoss	69.1	68.6	72.3	69.6	70.4	72.6	79.4	65.9	65.1	66.4	64.9	69.5

Table 9: Best reported performance on different kinship verification datasets.

Format	Dataset	Type	Method	Metric	Protocol	Score
Image	CornellKin [62]	4	Serraoui <i>et al.</i> 2022 [174]	Acc.	5-fold cross validation	93.8
	Family 101 [61]	4	Mukherjee <i>et al.</i> 2022 [139]	Acc.	5-fold cross validation	92.3
	KinFaceW-I [130]	4	Kohli <i>et al.</i> 2019 [100]	Acc.	5-fold cross validation	96.9
	KinFaceW-II [130]	4	Kohli <i>et al.</i> 2019 [100]	Acc.	5-fold cross validation	97.1
	TSKinFace [158]	4	Serraoui <i>et al.</i> 2022 [239]	Acc.	5-fold cross validation	90.7
	WVU [99]	7	Kohli <i>et al.</i> 2017 [99]	Acc.	5-fold cross validation	90.7
	IIITD [98]	7	Kohli <i>et al.</i> 2012 [98]	Acc.	5-fold cross validation	74.5
	FIW [166]	11	Robinson <i>et al.</i> 2016 [166]	Acc.	5-fold cross validation	71.0
Video	FIW(extended) [168]	11	Wang <i>et al.</i> 2018 [197]	Acc.	5-fold cross validation	69.5
	KFVW [226]	4	Yan <i>et al.</i> 2018 [226]	ERR.	5-fold cross validation	41.8
	KFVW [226]	4	Yan 2019 [225]	Acc.	5-fold cross validation	59.3
	TALKIN [35]	4	Wu <i>et al.</i> 2019 [216]	Acc.	5-fold cross validation	74.1
	UvA-Nemo Smile [49]	7	Dibeklioglu 2017 [48]	Acc.	5-fold cross validation	93.6
	KIVI [100]	7	Kohli <i>et al.</i> 2019 [100]	Acc.	5-fold cross validation	83.2

### Human evaluation

The evaluation of kinship verification by humans often occur in the domain of social analysis related topics [20, 42, 43, 64, 93, 95, 146, 154, 236]. The participants assessing the image pairs are usually divided by age, gender, race, career *etc.*. In a number of surveys, the participants generally tend to be specialists or students with basic psychological knowledge. In [93], 59 undergraduate students with an average age of 21.6 take part in the kinship verification test. These students all obtain partial credits in an introductory psychology course. In early research, human’s evaluation of kinship verification is questionnaire-based. Researchers show the different images to the participants without any labels. The participants write down the judgment of the kinship. In some experiments, the judgment time is recorded.


In recent years, machines are used to make experimental data more accurate. In [93], random stimuli appear on the screen. The participants need to judge the kinship between the pairs shown in the stimuli. The response time is limited to 20s. The response of the participants is also recorded. A degree of relatedness is finally recorded with the parameters of kinship assessment. In [125], the researchers used the Amazon Mechanical Turk service (MTurk) crowd-sourcing service to evaluate a set of kinship verification pairs. In these experiments, the MTurk participants are anonymous. Like the previous experiment, the pair of face images are displayed on the screen, and the participant’s answers are recorded by clicking the corresponding button. The final evaluation is the average score of all correct answers by all participants. Lopez *et al.* [125] show that for the dataset KinFaceW-I, KinFaceW-II, a human can reach a performance within a range of 75% to 85%. Both [93] and [125] show that humans are better, especially for M-D (mother and daughter) relationships.



**Task overview**  
We are interested in identifying if two people are part of the same family.

**Instructions**  
Look at the photos of the 2 people below and give your assessment if you think that they are related (i.e. part of the same family). Please keep in mind that the quality of your assessment will directly influence the quality of this research.

**Task**



Are these two people related (i.e. part of the same family):

☐ Yes

☐ No

Figure 21: Crowd-sourced human evaluation of kinship using Amazon Mechanical Turk, cited from Lopez et al. [125].

Table 10: Human evaluation for different datasets.

Year	Name	human	participants	description	train	score
2010	CornellKin	Fang et al. 2010 [62]	16	randomly select 20 pairs		67.19
2012	UB KinFace v2	Shao et al. 2011 [176]	20	40training sample, 40 test samples(20true)		time1: 53.17 time2: 56
2012	Family-Face	Xia et al. 2012 [221]	20	randomly select 32 pairs		56.88
2013	UvA-Nemo Smile	Lopez et al. 2018 [125]	304	MTurk+quality assurance		80.2
		Lopez et al. 2018 [125]	304	MTurk+quality assurance		78.6
2014	KinFaceW-I	Yan et al. 2014 [228] Lu et al. 2013 [130]	10	5 males and 5 females		71
		Lopez et al. 2018 [125]	304	MTurk+quality assurance		83.5
2014	KinFaceW-II	Yan et al. 2014 [228] Lu et al. 2013 [130]	10	5 males and 5 females		74
2015	TSKinFace	Qin et al. 2015 [158]	10	100 pairs A: parents -children 1:1 B parents -children 2:1	no no	74.62 79.55
			case1 75	406 samples, 11 categories, type-by-type basis	no	57.5
2016	FIW	Robinson et al. 2018 [168]	case2 110	406 samples, 11 categories, type unspecified	no	~57.5
				cropped, 20 pos,20 neg origin,		68.63
2017	KFVW	Yan et al. 2018 [226]	5m5f	20pos,20neg		73

### 2.7.2 Benchmark on the Nemo-Kinship Dataset

Different methods are re-implemented. All videos are pre-processed by a face detector and aligned with the same eye position.

#### Data post-processing

The pipeline of the post-processing of the Nemo-Kinship dataset is shown in Figure23. Firstly, we extract the members and divided them into 11 categories according to their

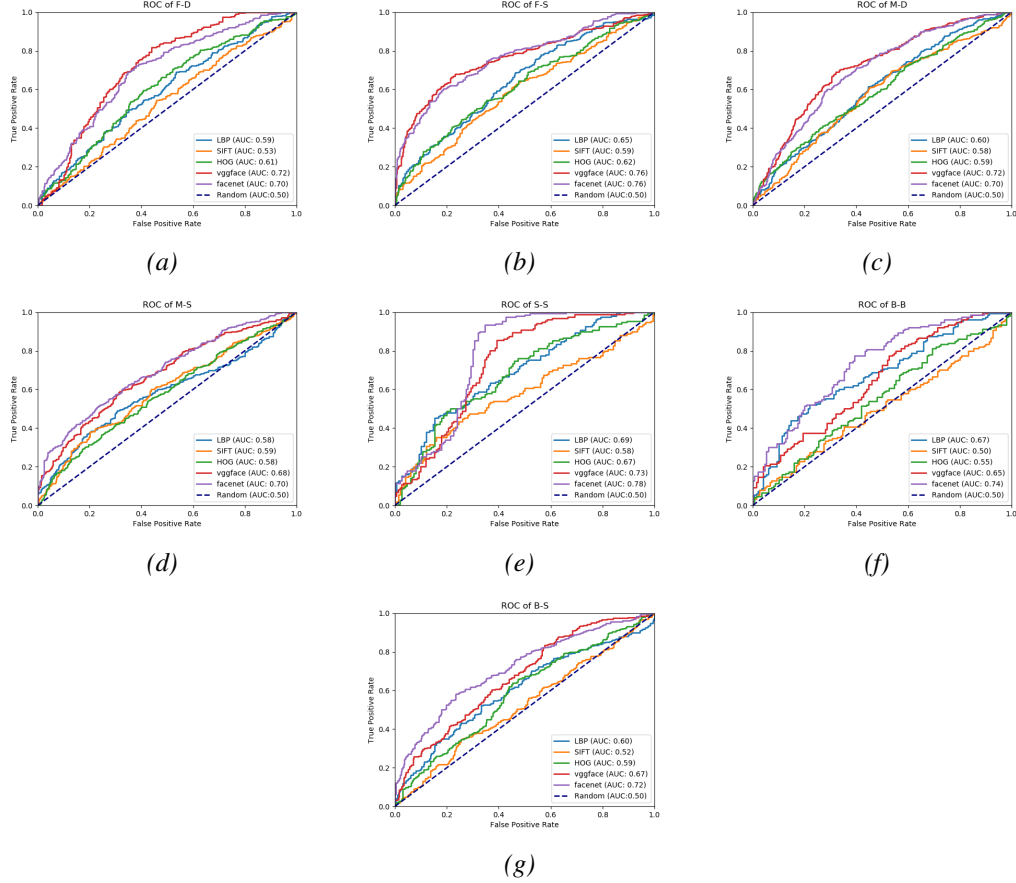


Figure 22: ROC curves based on different features.

kin-types. Since the number of samples, having the secondary kinship, is small, we only use seven kin relations as the testing relationship: M-D, M-S, F-D, F-S, B-B, B-S, and S-S. We extract one video from the Nemo-Kinship dataset with the "yes" answer. Secondly, we convert the video of each person into 100 frames. The faces are cropped into 160x160 pixels according to the bounding box of the detected face. Then, we align each image according to the landmarks. We adjust each face and fixed all the eye positions. Thirdly, all the family members are re-arranged into seven kinship-type folders. The entire dataset is trained and tested by 5-fold-cross-validation. Therefore, we generate a cross-validation list of five folders for each kinship for training and testing.

### Methods

Both image-based methods and video-based methods are selected. For image-based methods, NRML [130], CNN-points [240], Attention Networks [232], Sphreface-baseline [164, 165], and Vuvko [175] [175] are used. NRML is the traditional and widely-used metric learning method. CNN-points is the first deep learning method. Attention Network and Vuvko are more recent methods. Sphreface-baseline is the benchmark method for Recognizing Families In the Wild Data Challenge (RFIW) in 2020 and 2021. Vuvko reaches the state of the art results on the kinship verification track for RFIW2020. The performances of different methods are listed in Table 11. Among these methods, Vuvko shows the highest accuracy. Vuvko utilizes the information

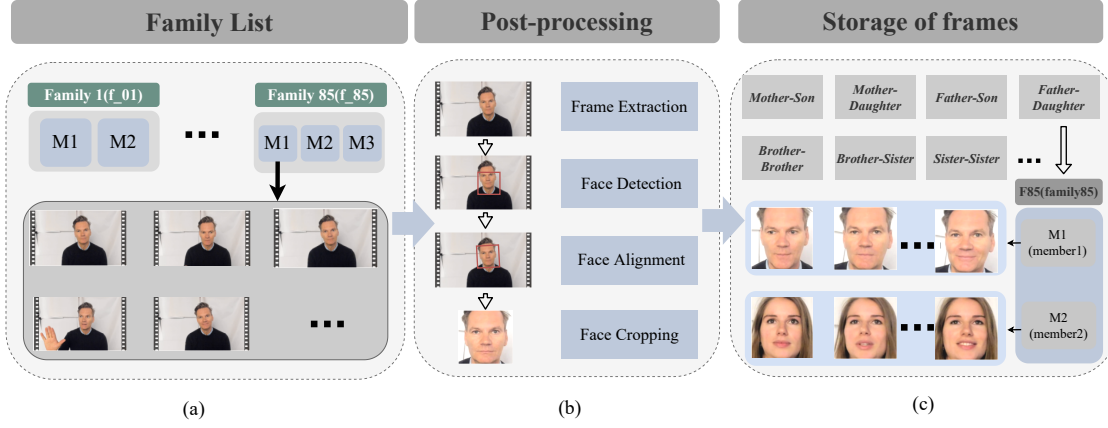


Figure 23: The pipeline of post-processing on the Nemo-kinship dataset. (a) The Nemo-kinship dataset is stored in 17 videos per family member considering the family tree.  $F_n$  represents the  $n$ th of the family group.  $M_n$  denotes the family members for each family. (b) All the videos are converted into frames. Every detected face is processed by the detection, alignment, and cropping procedure. (c) All family members are reconstructed into seven kinship folders. Each person contains 100 frames.

of the face recognition task and selects *arcface\_100\_v1* [47] as the backbone. The results show that face verification information helps to improve kinship verification. Comparing Attention Network (with masks) and Attention Network, it can be concluded that using masks improves the results. For video-based methods, the Deep+Shallow method proposed by Boutellaa *et al.* [18] is used. It combines deep features obtained by the convolutional network and spatio-temporal texture features. The results show an improvement for the brother-sister type.

### Feature representations

To study the influence of different features, SIFT, LBP, HOG, VGG-face, and Facenet are selected as basic descriptors. SIFT is one of the widely used feature descriptors in image recognition and classification. We follow [168] and [131]. The images are divided into  $16 \times 16$  blocks with a stride of 8. Then, the SIFT feature with  $128D$  is extracted from each block and concatenated together. The LBP [2] features are extracted following the implementation of [129]. The image is divided into  $16 \times 16$  non-overlapping blocks at first. The radius is set to 2, and the sampling number is set to 8. The extracted features are represented by  $256D$  histograms, forming a  $2096D$  ( $256 \times 16$ ) feature. Unlike traditional descriptors, VGG-face and Facenet are used as off-the-shelf face encoders following the settings of [168]. The similarities of image pairs based on different features are calculated by the cosine similarity with a certain threshold. The ROC curves of different features are shown in Figure 22. It shows that Facenet features and VGG-face features provide the best results. It also shows that test pairs with the same generation (Brother-brother, Sister-sister, and Brother-sister) obtain more distinguished features.

Table 11: Accuracy of existing methods on the Nemo-kinship dataset.

Method	Description	F-D	F-S	M-D	M-S	B-B	B-S	S-S	Average
Lu <i>et al.</i> 2012 [130]	NRML	0.6627	0.6209	0.5907	0.6321	0.7177	0.6198	0.7313	0.6209
	CNN-basic	0.5550	0.5462	0.5053	0.5790	0.5067	0.5393	0.5250	0.5366
Zhang <i>et al.</i> 2015 [240]	CNN-points	0.4965	0.6054	0.5075	0.5817	0.5597	0.5527	0.5067	0.5443
	Attention Network	0.5753	0.5396	0.5273	0.5585	0.4587	0.5523	0.5356	0.5353
Yan <i>et al.</i> 2019 [232]	Attention Network(with masks)	0.5792	0.5273	0.5672	0.5997	0.5290	0.5940	0.5700	0.5666
	Attention Network(Multi-inputs)	0.5288	0.5356	0.5137	0.5760	0.5077	0.5155	0.5520	0.5328
Shadrikov <i>et al.</i> 2020 [175]	Vuvko	0.7750	0.8488	0.7769	0.7335	0.8166	0.7607	0.7606	0.7715
Robinson <i>et al.</i> 2021 [164, 165]	Sphereface-baseline	0.5237	0.5407	0.5606	0.5548	0.5937	0.5803	0.5467	0.5572
Boutellaa <i>et al.</i> 2017 [18]	DEEP+Shallow	0.5833	0.5667	0.5756	0.5708	0.4667	0.7000	0.5333	0.5709

### 2.7.3 Discussion

The results of different methods on the public datasets and our newly proposed Nemo-Kinship dataset, show that the current methods (NRML, CNN-basic, CNN-points, Attention Network, Vuvko) provide better results on public datasets. This can be attributed to the fact that the Nemo-Kinship dataset contains more samples of children and adults. These samples show larger differences in appearance. Based on the results, Vuvko achieves the best performance. The features extracted by the combination of LBP and HOG are enhanced by metric learning during the training process. Due to overfitting, deep neural networks without pre-training do not show good results on the Nemo-kinship data set. Attention Network and Attention Network (with mask) show that attention to local features improves the performance. On the other hand, ROC curves of the differential feature extractors show that the deep features from the pre-trained network provide better features.

## 2.8 CONCLUSION

This survey provides a comprehensive review of public datasets and representative methods for kinship verification. Representative methods are categorized and compared based on their feature representations: (1) hand-crafted feature-based, (2) metric learning-based, and (3) deep learning-based. Also, this review studies current kinship challenges according to intrinsic factors (face *i.e.*, differences in facial appearance) and extrinsic factors (acquisition *i.e.*, varying imaging conditions). New promising directions are discussed based on current advances in kinship research. Open-set kinship verification and debiasing kinship verification are largely ignored so far. They are promising for the kinship verification task in the future. Through the analysis of current kinship verification datasets, we believe that there is still a need for more kinship datasets for specific problems. More video-based kinship datasets are in demand. Therefore, a new video dataset is presented as a benchmark for a child-adult-based kinship verification task. This dataset consists of 248 subjects from 85 families. It contains age, gender, and audio information. This benchmark is used to systematically test and analyze current state-of-the-art methods.

---

## KINSHIP IDENTIFICATION THROUGH JOINT LEARNING

---

### 3.1 INTRODUCTION

Kinship is the relationship between people who are biologically related with overlapping genes [129, 131], such as parent-children, sibling-sibling, and grandparent-grandchildren [6, 166, 170, 217]. Image-based kinship identification is used in a variety of applications including missing children searching [217], family album organization, forensic investigation [170], automatic image annotation [129], social media analysis [20, 43, 236], social behavior analysis [64, 93, 146, 240], historical and genealogical research [43, 95], and crime scene investigation [100].

While kinship verification is a well-explored task, identifying whether or not persons are kin, kinship identification, which is the task to further identify the particular type of kinship, has been largely ignored so far. Existing kinship verification methods usually train and test each type of kinship model independently [166, 197, 217] and hence do not fully exploit the complementary information among different kin types. Moreover, existing datasets have unrealistic positive-negative sample distributions. This leads to significant limitations in real world applications. When conducting kinship identification, since there is no prior knowledge of the distribution of images, all independently trained models are used to determine the kinship type of a specific image pair. Figure. 24 shows an example of providing an image pair to four individually trained verification networks based on a recent state-of-the-art method by Yan *et al.* [232]. The network generates contradictory outputs showing that the test subjects are simultaneously father-daughter, father-son, mother-son and mother-daughter.

In this chapter, a new identification method is proposed to learn the identification and verification labels jointly i.e. combining the kinship identification and verification tasks. Specifically, all kinship-type verification models are ensembled by combining the binary output of each verification model to form a multi-class output while training. The binary and multi-class models are leveraged in a multi-task-learning way during the training process to enhance generalization capabilities. Also, we propose a baseline multi-classification neural network for comparison.

We test our proposed kinship identification method on the KinFaceW-I and KinFaceW-II datasets and demonstrate state-of-the-art performance for kinship identification. We

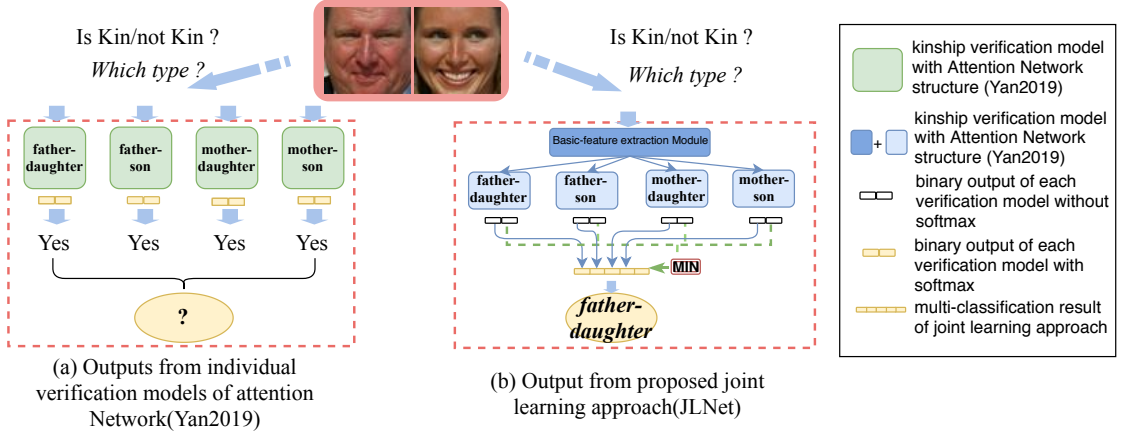


Figure 24: Identification of kinship relationships using verification ensembles. (a) Existing verification networks are trained independently resulting in contradictory outputs. (b) The output of our proposed joint training

also show that the proposed method significantly improves the performance of kinship verification when trained on the same unbiased dataset.

To summarize, the contributions of our work are:

- We propose a theoretical analysis in metric space of relationships between kinship identification and kinship verification.
- We propose a joint learnt network that simultaneously optimizes the performance of kinship verification and kinship identification.
- The proposed method outperforms existing methods for both kinship identification and unbiased kinship verification.

### 3.2 RELATED WORK

**KINSHIP VERIFICATION** Fang *et al.* [62] are the first to use handcrafted feature descriptors for kinship verification. Later, Xia *et al.* collected a new dataset with young and old parent images to utilize the intermediate distribution using transfer learning [219, 220]. Lu *et al.* [131, 249] propose a series of metric learning methods. Other handcrafted feature-based methods can be found in [49, 61, 129, 204, 219, 227, 229, 251]. Deep learning-based methods [232, 240] exploits the advantages of deep feature representations by using pre-trained neural networks in an off-the-shelf way. Zhang *et al.* are the first to use deep convolutional neural networks [240], and Yan *et al.* [232] are the first to add attention mechanisms in deep learning networks for kinship verification. In recent years, there is a trend to combine different features from both traditional descriptors [227, 249] and deep neural networks [21, 88, 184] to generate better representations [18]. (m)DML [52, 198] combines auto-encoders with metric learning. However, these methods focus on specific types of kinship and train and test on the same kinship types separately, which may not be feasible in real-world scenarios.

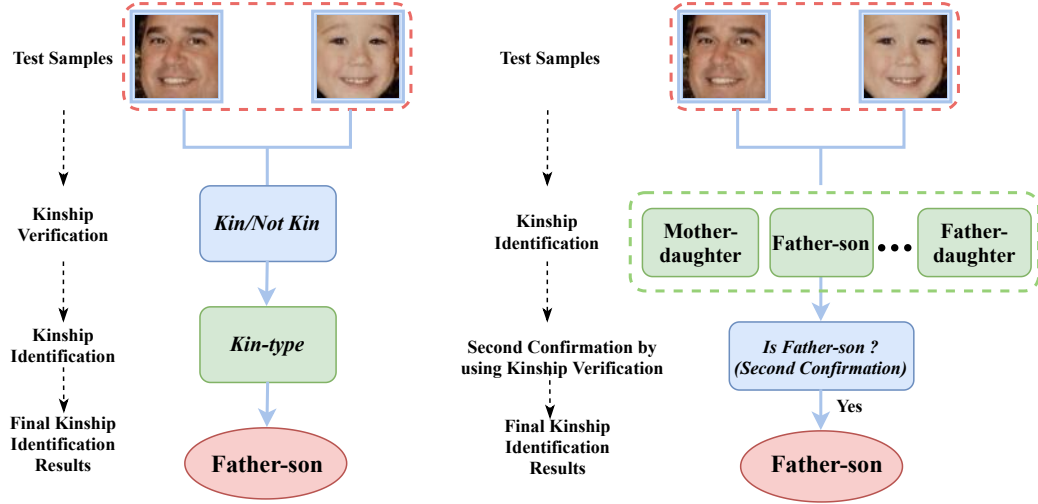
**KINSHIP IDENTIFICATION** Different from kinship verification, kinship identification attracted less attention [6]. [6, 166] only slightly deal with kinship identification. Guo *et al.* [78] propose a pairwise kinship identification method using a multi-class linear logistic regressor. The method uses graph information from one image with multi inputs. The paper is based on "kinship recognition" and uses a strong assumption that all the data is processed by a perfect kinship verification algorithm. Since there is not sufficient data with family annotations, the method is limited by using multi-input labels. In contrast, our method handles negative pairs and focuses on pair-wise kinship identification. For example, in the context of searching for missing children, we need to handle each potential pair online and find the most likely pair for specific kinship types. In this case, we need to filter the online data and test the most likely data after filtering. As for the family photo arrangement or social media analysis, the aim is to understand the relationships between persons in a picture. There are usually many faces and different kinship relations in a family picture. Hence, the goal is to verify the most likely pairs among negative pairs. Previous methods are not able to cope with this scenario. Figure 25 shows that kinship verification is closely related to kinship identification. As a consequence, we propose a new approach by jointly learning all independent models with kinship verification and identification information.

### 3.3 KINSHIP IDENTIFICATION THROUGH JOINT LEARNING WITH KINSHIP VERIFICATION

In this section, we first introduce the three types of relationship understanding: kinship verification, kinship identification, and kinship classification. Based on this, we introduce the current challenge on kinship identification. Finally, we introduce the concept of conducting kinship identification by using a joint learning strategy between kinship identification and kinship verification.

#### 3.3.1 *Definition of Kinship Verification, Kinship Identification and Kinship Classification*

Kinship recognition is the general task of kinship analysis based on visual information. There are mainly three sub-tasks [6, 166]: kinship verification, kinship identification, and kinship classification (e.g. family recognition). The goal of kinship verification is to authenticate the relationship between image pairs of persons by determining whether they are blood-related or not. Kinship identification aims at determining the type of kinship relation between persons. Kinship classification [166, 217] is the recognition of the family to which a person belongs to. Figure 25 illustrates the relationship between these tasks. This chapter focuses on kinship identification, which is an important but not well-explored topic. Unlike other kinship recognition methods [32, 78, 194, 203], which take images of multiple people as input to predict the relationships between them, the kinship identification task targets at classifying the kin-type of image pairs (negative pairs also included).



(a) A common process of kinship analysis

(b) Detailed steps of kinship identification

Figure 25: Flowchart of the relation between kinship verification and kinship identification. (a) Kinship verification is used as a preliminary process for kinship identification. (b) The kinship identification process can be divided into two steps: kinship identification and kinship verification on a specific type.

### 3.3.2 Relationship between Kinship Verification and Kinship Identification and the Limitation of Existing Methods

#### Relation between the Two Tasks

In the literature, kinship verification and identification are two tasks which are studied separately but are closely related. When analyzing the kinship relation between persons, verification is usually applied first to determine whether these persons are kin or not. Then, the kinship type is defined. Figure 25.a shows the common process of kinship analysis, where kinship verification is used as a preliminary process for kinship identification. Furthermore, kinship identification can be divided into two steps, as shown in Figure 25.b. In the first step, the images are preliminarily classified by the kinship identification model. Then, the classified images are sent to the corresponding verification model. Due to the differences of inherited features among different kin-type images, the kinship verification model provides a better representation than a general kinship identification model. On the other hand, since the kinship identification process filters out irrelevant samples, it provides a consistent and similar feature distribution for kinship verification modelling. In this way, kinship verification and identification are two complementary processes, and can benefit from each other.

#### Representation of Kinship Relationships in Metric Feature Space and Limitation of Existing Methods

In the literature, metric learning is a popular approach for kinship verification. Ideally, the learnt metric space represents kinship likeness for smaller distances. However, existing kinship verification models only consider specific kinship types and ignore the influence of other types.



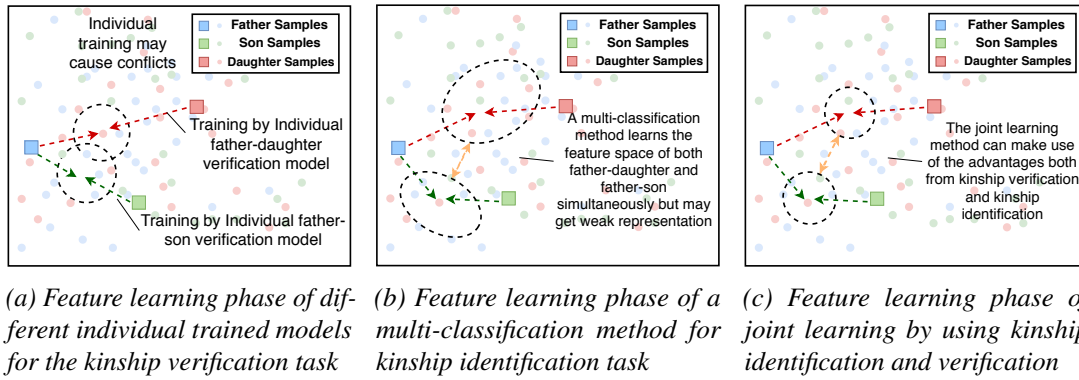


Figure 26: Feature space of models during training. Similar feature shapes indicate that the samples are from the same family. Joint learning better represents the context between different kinship relationships. Small circles are used to represent focused samples in feature space.

As shown in Figure 26a, when the father-daughter verification model is being trained, the features of father and daughter samples will be congregated during the training process and the negative daughter images will be pulled apart. However, due to the negative samples of father-son pairs, which are not included in the training data, the features of son images are less affected by the training process pulling father-son images apart. A narrow-down training of kinship verification can improve the representation of each sample within a specific kin-type. However, since the model does not thoroughly learn other types of negative samples, the separate trained models can easily conflict with each other resulting in ambiguous results. In contrast, a multi-classification method not only considers different types of images but also the interaction between different types. As shown in Figure 26b, the son features will be learned as negative features for the father-daughter feature space, whereas the features of daughters will be considered as negative features for the father-son space. The yellow arrows in Figure 26b indicate negative samples which will be separated from the matched feature space. A multi-classification method may obtain a weaker representation for a specific kin-type because of the large difference of inherited features among different kin-type images. A joint learning method has the advantage of the generalization of multi-class training and the representation of individual verification models. Hence, identification methods based on joint learning not only repulse negative pairs of different kinship types but also push the potential negative images to the target feature space, which is illustrated in Figure 26c.

#### Real World Kinship Distribution and Dataset Bias

Note that the proportion of positive and negative samples is highly unbalanced for existing kinship verification datasets in the real world. This unbalanced distribution has a negative impact on different applications. Take the online family picture organization application for example. The problem is to determine the matched pairs of images for a specific kinship relationship when the number of kin-related samples only contains a small portion of the entire dataset. Another example is that, when searching for missing children, to retrieve a picture that looks the most like the son of the parents in which the majority of these samples are negative samples.

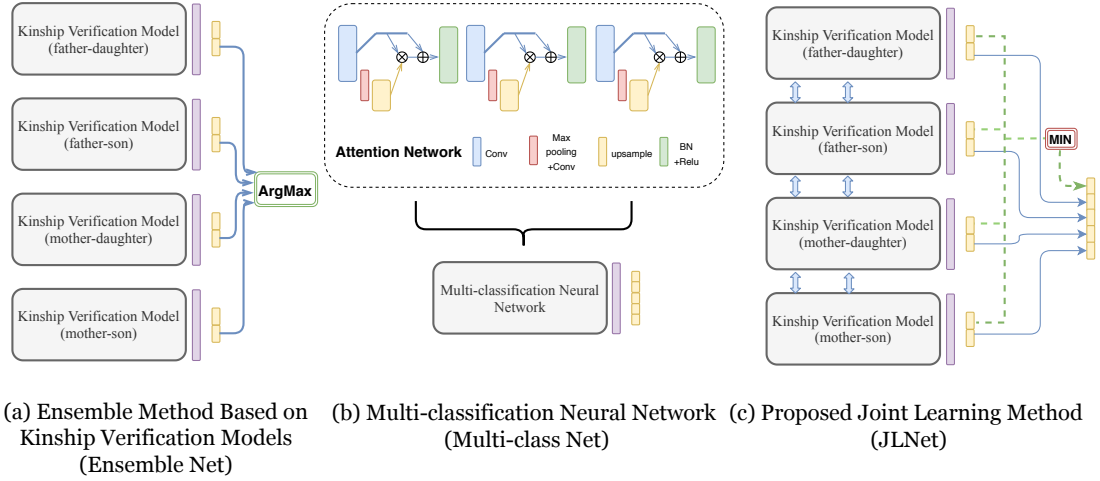


Figure 27: Structure of the approaches using four relationships as an example.

### 3.4 JOINT LEARNING OF KINSHIP IDENTIFICATION AND KINSHIP VERIFICATION

We propose a joint learning network (JLNet) based on the learning strategy shown in Figure 27 aiming to utilize the representation capability of kinship verification models as well as making use of the advantages of multi class classification. This approach consists of two major steps: the combination of different types of images and joint learning.

The main ideas of the approach are summarized as follows:

1. We utilize all different kin-types of image pairs to train each kinship model, not based on a specific type.
2. Different models are trained jointly to differentiate negative kinship feature pairs from the matched model and to merge positive pairs as much as possible.

Note that naively using a single classification network (Figure 27.a) or naively combining multiple verification networks (Figure 27.b) are not suitable approaches. As described above, our network (Figure 27.c) utilizes the advantage of both tasks. Without loss of generality, we outline our approach for four relationships: father-daughter (F-D), father-son (F-S), mother-daughter (M-D), mother-son (M-S).

#### 3.4.1 Architecture of the Proposed Joint Learning Network (JLNet)

The new Joint Learning Network (JLNet) is illustrated in Figure 28. The structure of JLNet consists of two parts: the individual Verification Module and the Joint Identification Module.

##### *Individual Kinship Verification Module*

As shown in Figure 27.c, each Individual Kinship Verification Module is defined as a binary classification problem. Let  $\mathbf{S} = \{(I_{p_i}^\alpha, I_{c_j}^\beta), i, j = 1, 2, \dots, N, \alpha = 1, 2, 3, 4, \beta = 1, 2, 3, 4\}$  be the training set of  $N$  pairs of images. And  $\alpha \in \{1, 2, 3, 4\}$  and  $\beta \in \{1, 2, 3, 4\}$  correspond

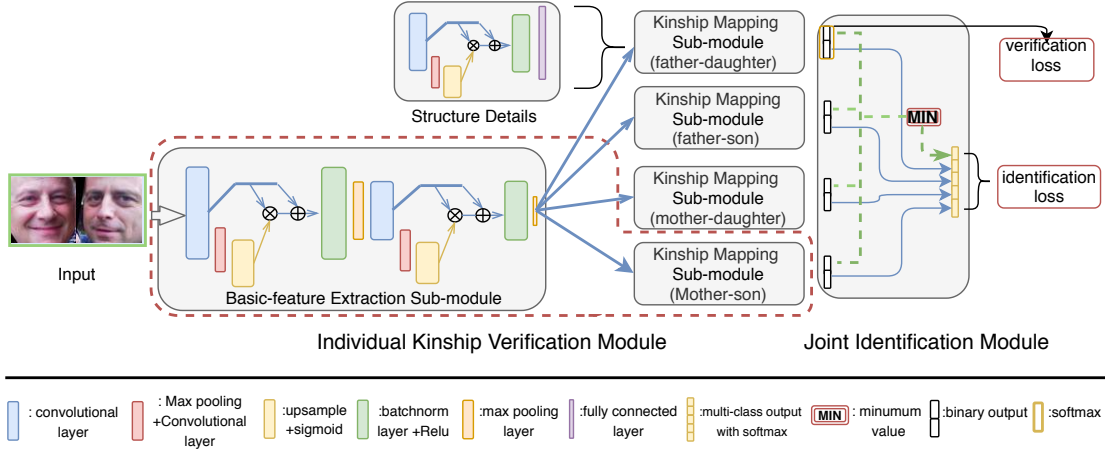


Figure 28: Architecture of our Joint Learning Network (JLNet)

to the following kinship types: father-daughter, father-son, mother-daughter, mother-son respectively. Then, the Individual Verification Module is defined by:

$$\hat{y} = \mathcal{D}_{\theta}^n(I_{p_i}^{\alpha}, I_{c_j}^{\beta}), \quad (12)$$

where  $I_{p_i}^{\alpha} \in \mathbb{R}^{H \times W \times 3}$  is  $i$ -th parent image from  $\alpha$  type data set and  $I_{c_j}^{\beta} \in \mathbb{R}^{H \times W \times 3}$  is the  $j$ -th child image from  $\beta$  type data set. The output  $\hat{y}$  of each Individual kinship verification Module is a  $1 \times 2$  vector. An Attention Network [232] is used as the basic architecture for each Individual Kinship Verification Module. As shown in Figure 28, the Attention Network uses a bottom-up top-down structure and consists of three attention stages. Each stage consists of one attention module and one residual structure. To exploit the shared information between the complimentary tasks, the parameters of the two stages of the Attention Network are shared to learn low-level and mid-level features from the input images. This forms the Basic-feature Extraction Sub-module. This Basic-feature Extraction Sub-module extracts the basic, generic facial features. Then, high-level features are extracted: four separate branches are added after the last layer (a max pool layer) of the Basic-feature Extraction Sub-module. Each branch focuses on one specific kin-type separately, resulting in four Kinship Mapping Sub-modules. Each of this sub-Module obtains the third stage of the Attention Network and focuses on different kinship types.

#### Joint Identification Module

The binary outputs of each Individual Kinship Verification Module are ensembled. The binary output is described in Eq. 12. The multiple output  $\hat{O}$  of the kinship identification module is defined by:

$$\hat{O}_m = \begin{cases} \min_{n \in \{1, 2, 3, 4\}} \mathcal{D}_{\theta}^n(I_{p_i}^{\alpha}, I_{c_j}^{\beta})_{z=1}, & \text{if } m = 0 \\ \mathcal{D}_{\theta}^n(I_{p_i}^{\alpha}, I_{c_j}^{\beta})_{z=2} & \text{if } m \neq 0 \end{cases}, \quad (13)$$

where  $m \in \{0, 1, 2, 3, 4\}$  represents the  $m$ -th item of vector  $\hat{O}$  and  $z$  represents  $z$ -th item of the output vector of  $\mathcal{D}_\theta^n$ . The output class  $C$  is defined by:

$$C = \arg \max_{m \in \{0,1,2,3,4\}} \sigma(\hat{O})_m, \quad (14)$$

where  $\sigma(\cdot)$  is the softmax function.

During the training, the Weighted Cross Entropy loss is used for both kinship verification and identification:

$$\mathcal{L} = - \sum_{i=1}^n w_n \log(\sigma(\cdot)_n), \quad (15)$$

where  $n$  is the class label of the kinship verification or identification output and  $\sigma(\cdot)_n$  is the  $n$ -th output of the softmax function. The loss of the joint learning model is given by a weighted summation of the kinship verification loss (from binary outputs) and the kinship identification loss (from multiple outputs):

$$\mathcal{L} = \sum_{i=1}^4 \lambda_i \mathcal{L}_{kvi} + \lambda_5 \mathcal{L}_{kI}, \quad (16)$$

where  $\mathcal{L}_{kI}$  is the Weighted Cross Entropy loss of the kinship identification output given by Eq. 15 and  $\lambda_i$  is the  $i$ -th weight of each loss.

### 3.4.2 Comparative Methods

#### *Ensemble Method based on Kinship Verification Models (Ensemble Net)*

Figure 27.a shows the structure of the Ensemble Method based on Kinship Verification Models (Ensemble Net). The Individual Kinship Verification Modules of the Ensemble Net have the same structure as JLNet. While testing, the Ensemble Net feeds the images into four kinship verification models simultaneously and ensembles four binary outputs. The output class  $C$  is defined by:

$$C = \begin{cases} 0 & \text{if } \max_n \sigma(\mathcal{D}_\theta^n(I_{p_i}^\alpha, I_{c_j}^\beta))_{z=2} < 0.5 \\ \arg \max_n \sigma(\mathcal{D}_\theta^n(I_{p_i}^\alpha, I_{c_j}^\beta))_{z=2}, & \text{otherwise} \end{cases}, \quad (17)$$

where  $I_{p_i}^\alpha$  is the  $i$ -th parent image from  $\alpha$  type data set and  $I_{c_j}^\beta$  is the  $j$ -th child image from  $\beta$  type data set.

#### *Multi-Classification Neural Network (Multi-class Net)*

The structure of the Multi-Classification Neural Network (Multi-class Net) is shown in Figure 27.b. Similar to the Ensemble Net, Multi-class has the same backbone with the Individual Kinship Verification Module of JLNet. The Multi-class Net handles the kinship identification task as a multiple classification problem:

$$\hat{y} = \mathcal{D}_\theta(I_{p_i}^\alpha, I_{c_j}^\beta), \quad (18)$$

where  $\mathbf{S} = \{(I_{p_i}^\alpha, I_{c_j}^\beta), i, j = 1, 2, \dots, N, \alpha = 1, 2, 3, 4, \beta = 1, 2, 3, 4\}$  and the output  $\hat{y}$  is a  $1 \times 5$  vector.

### 3.5 EXPERIMENTS

#### 3.5.1 *Unbias Dataset for Training and Testing*

Three types of benchmark datasets are generated from the KinFaceW-I and KinFaceW-II datasets [129, 131] consisting of four kinship types: father-daughter (F-D), father-son (F-S), mother-daughter (M-D), mother-son (M-S). To conduct the experiment on unbiased datasets, we re-balance the KinFaceW-I and KinFaceW-II datasets into three different benchmark datasets as follows:

1. *Independent Kin-type Image Set*: This dataset has four independent subsets, where each subset contains one specific kinship type. This dataset simulates a dataset obtained by an ideal kinship classifier. The split of this image set is the same as KinFaceW-I or KinFaceW-II. The positive samples are the parent-children pairs with the same type of kinship. The negative samples are the pairs of unrelated parents and children within the same kin-type distribution. The positive and negative ratio is 1 : 1.
2. *Mixed Kin-Type Image Set*: This dataset combines four different kin-type images taken from the KinFaceW-I or KinFaceW-II datasets resulting in the type ratio (father-daughter: father-son: mother-daughter: mother-son: negative pairs) to be 1 : 1 : 1 : 1 : 4. This image set is used for both training and testing. Image pairs with kinship relations are denoted as positive samples. Negative samples are random image pairs without kinship relation but within the same type of distribution.
3. *Real-Scenario Kin-Type Image Set*: This dataset simulates the data distribution for real-world scenarios (e.g. retrieval of missing children). All the images in the KinFaceW-I or KinFaceW-II datasets are paired one by one, which leads to a highly unbalanced positive-negative rate. Taking KinFaceW-II as an example, in each cross-validation, there will be 400 images (200 positive pairs) to be tested. All these images are paired one by one. The ratio of positive and negative pairs is 1 : 398.

#### 3.5.2 *Experimental Design*

All methods are trained on the Mixed Kin-Type Image Set. The dataset is divided into 5-folds and verified by a 5-cross validation. We use the same data augmentation for all methods. The data is augmented by randomly changing the brightness, contrast, and saturation of the image. Random grayscale variations, horizontal flipping, perspective changes, and resizing and cropping are also included. All images have the same size  $64 \times 64 \times 3$ , and the batch size is set to be 64.

### *Proposed Joint Learning Method (JLNet)*

The training scheme of JLNet is divided into two phases. The first one is to train the network parameters for the four models independently. The weighted cross entropy is used for updating and the weight list is set to be  $[0.25, 8]$  for each verification output. The second phase is to update network parameters jointly by using both binary and multiple-outputs. The weight matrix of the cross-entropy of the multiple outputs is set to  $[0.18, 2, 2, 2, 2]$ , and  $\lambda_i$  of the total loss is  $1 : 1 : 1 : 1 : 10$  respectively. Adam is used as optimizer and the learning rate is set to  $10^{-4}$ . Since there is no public code available for the attention network, we re-implemented the attention network from scratch. During testing of the kinship verification of each individual kin-type, the binary output of the matched Individual Kinship Verification Module is taken as the final result. During testing of the kinship identification task, both the binary outputs (for kinship verification) and multiple outputs (for kinship identification) are used. A combined result based on the confidence of these two types of outputs are taken as the final result.

### ABLATION STUDY

- *Joint Learning without Backpropagation of Multiple Outputs (JLNet<sup>†</sup>)*: To assess the performance of additional multi-classification outputs, the structure of JLNet<sup>†</sup> is kept the same as JLNet. Further, JLNet<sup>†</sup> is trained in the same way as JLNet, but without using multiple output results for parameter updating.
- *Joint Learning using Multiple Outputs for Kinship Identification (JLNet<sup>‡</sup>)*: We use the trained model of JLNet directly but only the multiple output is taken as the final result during testing.

### *Experiments and Comparison*

**ENSEMBLE NET** For Ensemble Net, we provide two ways to train the models:

- *Ensemble Net\**: Each verification model is trained separately on the Independent Verification Image Set, which is the same as [232]. This means that each independent kinship verification module is only trained on matched data.
- *Ensemble Net*: Each verification model is trained on the Mixed-Type Image Set, which is the same as the training data of JLNet and Multi-class Net. Adam is used and the learning rate was set to be  $10^{-4}$ . The weights of the cross entropy are 0.25, 8.

**MULTI-CLASS NET** Also for the Multi-Class Net, Adam is used as an optimizer. The learning rate is again  $10^{-4}$ . A weight list of  $[0.1, 1, 1, 1]$  is used for the weighted Cross Entropy loss.

### 3.5.3 *Results & Evaluation*

The methods are evaluated on the different datasets. Five-cross validation is used as the evaluation protocol. As a reminder, Ensemble Net\* is trained on the Independent

Kin-Type Kinship Image Set, JLNet<sup>†</sup> is trained without Backpropagation of Multiple Outputs, and JLNet<sup>‡</sup> uses multiple outputs as the final result. The results are shown in Table 12-16.

#### Results for the Independent Kin-Type Image Set

Table 12: The accuracy of different methods through 5-fold cross-validation on the Independent Kin-Type Image Set.

Methods	KinFaceW-I					KinFaceW-II				
	F-D	F-S	M-D	M-S	Mean	F-D	F-S	M-D	M-S	Mean
Ensemble Net*	0.7017	0.7506	0.7410	0.615	<b>0.7021</b>	0.746	0.7440	0.7520	0.7320	<b>0.7435</b>
Multi-class Net	0.6463	0.6797	0.6650	0.5770	0.6420	0.5880	0.6240	0.6200	0.5920	0.6060
Ensemble Net	0.6425	0.6321	0.6382	0.577	0.6224	0.6060	0.6000	0.5860	0.6260	0.6045
JLNet <sup>†</sup>	0.6534	0.6991	0.6539	0.5772	0.6459	0.6160	0.6100	0.600	0.6500	0.6190
JLNet	0.6608	0.7309	0.7207	0.5897	<b>0.6755</b>	0.6800	0.7140	0.6860	0.7060	<b>0.6965</b>

Table 13: F1 scores of different methods through 5-fold cross-validation on Independent Kin-Type Image Set

Methods	KinFaceW-I					KinFaceW-II				
	F-D	F-S	M-D	M-S	Mean	F-D	F-S	M-D	M-S	Mean
Ensemble Net*	0.6915	0.7472	0.7566	0.6648	<b>0.7150</b>	0.7671	0.7589	0.7690	0.7607	<b>0.7639</b>
Multi-class Net	0.6084	0.6563	0.6767	0.5766	0.6295	0.5629	0.6000	0.6143	0.5062	0.5709
Ensemble Net	0.6639	0.6737	0.6735	0.6083	<b>0.6548</b>	0.6213	0.6439	0.6051	0.6399	0.6276
JLNet <sup>†</sup>	0.6301	0.6952	0.6496	0.5816	0.6391	0.6396	0.6166	0.6061	0.6191	0.6203
JLNet	0.6320	0.7087	0.7052	0.5657	0.6529	0.6585	0.7211	0.6939	0.6847	<b>0.6896</b>

Table 12 shows the verification results for the different methods based on the Independent Kin-Type Kinship Image Set. For this image set, accuracy and F1 scores are used to evaluate the performance of kinship verification. All methods are trained on the Mixed Kin-type Image Set except for ensemble Net\*. The results show that when trained on the same dataset, JLNet outperforms all other approaches. When tested on the KinFaceW-II dataset, JLNet outperforms Multi-Class Net with 9% and Ensemble Net by 9.2% on average accuracy. Considering the F1 score, JLNet outperforms Multi-Class Net with 11.9% and Ensemble Net with 6.2% on average. When comparing JLNet<sup>†</sup> and JLNet, it is shown that additional multi-outputs improve the results of the ensembled models. When compared with Ensemble Net, the accuracy of JLNet is lower than Ensemble Net. One of the reasons is that each of the verification module of Ensemble Net is trained on one specific dataset. This may result in overfitting. JLNet provides better generalization than Ensemble Net\*, as shown in the next section.

#### Results on Mixed Kin-Type Kinship Image Sets

Table 14 shows the results of macro F1 scores and accuracy for the kinship identification task using the Mixed Kin-Type Kinship Image Set. The results show that the performance of JLNet outperforms the ensemble and multi-class net methods. Moreover, macro F1 scores show that JLNet(full) outperforms Ensemble Net\* with 22.7% on KinFaceW-I and with 25.0% on KinFaceW-II. Moreover, JLNet(full) outperforms Ensemble Net\* with

Table 14: Macro F1 score and accuracy of kinship identification for the Mixed Kin-Type Kinship Image Set

Methods	KinFaceW-I		KinFaceW-II	
	macro F1	Accuracy	macro F1	Accuracy
Ensemble Net*	0.3240	0.3723	0.2846	0.3319
Multi-class Net	0.5291	0.5494	0.4861	0.5225
Ensemble Net	0.4837	0.4887	0.4464	0.4564
JLNet <sup>†</sup>	0.5155	0.5487	0.4648	0.4875
JLNet <sup>‡</sup>	<b>0.5507</b>	0.5880	0.5285	0.5535
JLNet(full)	0.5506	<b>0.5993</b>	<b>0.5343</b>	<b>0.5790</b>

22.7% on KinFaceW-I and 24.7% on KinFaceW-II. As shown in Figure 29, Ensemble Net\* may lead to indecisive results. The independently trained verification models can lead to overfitting and results in weak generalization capabilities. JLNet obtained the highest performance. In Figure 29, it is shown that the joint learning method provides indecisive results. To this end, the joint learning method JLNet(full) obtains the best performance for kinship identification on the Mixed Kin-type Kinship Image Set.



Figure 29: Confusion matrix for different experiments on the Mixed Kin-Type Image Set using the KinFaceW-I dataset. Negative samples are excluded)

### Results on Real Scenario Sample Set

Tables 15 and 16 show the results of the F10 score and accuracy for the kinship identification task in a real-world scenario. We focus more on recall than precision, so the F10 score is used to emphasize on the recall rate. The results show that JLNet(full) obtains the best performance on both KinFaceW-I-based Real-Scenario data and KinFaceW-II-based Real-Scenario data. The results show that the JLNet(full) outperforms all the other approaches for both KinFaceW-I and KinFaceW-II. From the confusion matrix in Figure 30, it is interesting to note that father-son and mother-daughter relations are more distinguishable than other kin-types. We argue that the manifold of pairs with the same gender is easier to be learned.



Table 15: F10 score and accuracy for different methods on the Real-Scenario Set using KinFaceW-I dataset. F10 (all) represents the average of F10 scores for all different labels (the negative label is also included)

methods	KinFaceW-I						
	F-D	F-S	M-D	M-S	mean	F10(all)	Accuracy
Ensemble Net*	0.0886	0.1179	0.1236	0.1003	0.1076	0.1830	0.4807
Multi-class Net	0.1548	0.2951	0.3047	0.1539	0.2271	0.2947	0.5618
Ensemble Net	0.1508	0.2791	0.2740	0.1378	0.2104	0.2596	0.4537
JLNet <sup>†</sup>	0.1522	0.2966	0.2937	0.1569	0.2249	0.2985	0.5901
JLNet <sup>‡</sup>	<b>0.1742</b>	0.3235	0.3123	0.1620	0.2430	0.3287	0.6681
JLNet(full)	0.1715	<b>0.3241</b>	<b>0.3198</b>	<b>0.1669</b>	<b>0.2456</b>	<b>0.3459</b>	<b>0.7439</b>

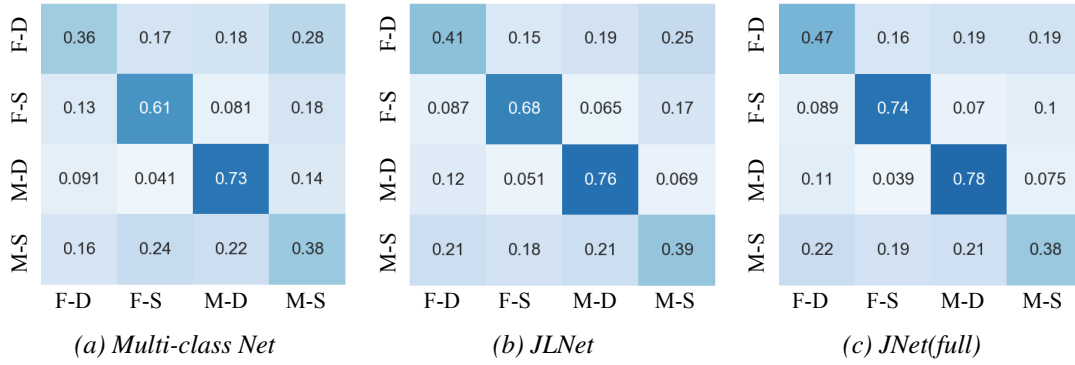


Figure 30: Confusion matrix for different experiments on the Real-Scenario Image set using the KinFaceW-I dataset. Negative samples are excluded)

Table 16: F10 score and accuracy for different methods on the Real-Scenario Set using KinFaceW-II dataset. F10 (all) represents the average of F10 scores for all different labels (the negative label is also included)

methods	KinFaceW-II						
	F-D	F-S	M-D	M-S	mean	F10(all)	Accuracy
Ensemble Net*	0.0469	0.0713	0.0726	0.0904	0.0703	0.1498	0.4647
Multi-class Net	0.1468	0.1972	0.1853	0.1076	0.1592	0.2528	0.6240
Ensemble Net	0.1399	0.1681	0.1496	0.0900	0.1369	0.2075	0.4874
JLNet <sup>†</sup>	0.1413	0.1757	0.1624	0.0962	0.1439	0.2303	0.5730
JLNet <sup>‡</sup>	0.1620	0.2133	0.2127	0.1225	0.1776	0.2735	0.6547
JLNet(full)	<b>0.1867</b>	<b>0.2134</b>	<b>0.2296</b>	<b>0.1296</b>	<b>0.1898</b>	<b>0.3003</b>	<b>0.7398</b>

### 3.6 CONCLUSION

In this chapter, we presented a new approach for kinship identification by joint learning. Experimental results show that joint learning with kinship verification and identification improves the performance of kinship identification. To our knowledge, this is the first approach to handle the kinship identification tasks by using deep neural networks jointly. Since this method is not restricted to any neural network, a better architecture can further improve the performance for kinship identification.



---

## IDENTITY INVARIANT AGE TRANSFER FOR KINSHIP VERIFICATION OF CHILD-ADULT IMAGES

---

### 4.1 INTRODUCTION

Image based kinship verification is an important computer vision task with different applications such as social media analysis [236], social behavior analysis, and historical and genealogical research [95]. Different methods focussing on (1) unconstrained conditions [134], (2) large scale datasets [167] and (3) multi-kinship types [194] have been proposed. However, the kinship verification problem between children and adults has been largely ignored. This is a tremendous problem as there are many applications involved including the search for missing children, the adoption of children and the creation of family albums [227].

Kinship verification for child-adult pairs is difficult because there are large appearance differences between children and their parents. Furthermore, as shown in Figure 31.a, there may exist a higher similarity in facial appearance in feature space for non-blood-related child-adult pairs than blood-related child-adult pairs. This may negatively influence the feature representation and may confuse the verification model. To mitigate this discrepancy, we propose a method to transform children’s face images into adulthood faces while maintaining their identity information. As depicted in Figure 31.c, the latent features of children shift towards the feature space of their parents. In this way, the difference in facial appearance between children and their parents is mitigated.

To this end, we propose a Children-Adult-Transferring (CAT) Module to extract identity and age-related features by using pre-trained information from an Identity-preserved Aging Generator. The face image of a child is transferred to an older age range while the parent’s face image is kept in a similar age range. Then, the extracted features are used as input for a Kinship Mapping Module, which maps the extracted features to a kinship-related manifold. Finally, the mapped kinship-related features are processed by a Neighborhood Repulsed Metric Learning (NRML) [131] for kinship verification.

In Figure 31, our proposed method is explained. Most of the publicly available kinship datasets contain relatively unbalanced age distributions, where the child-adult pairs usually occupy only a small portion of the dataset. Therefore, we created the Nemo-Kinship-Children dataset by collecting children-adult images. The main contribution of this chapter can be summarized as follows:

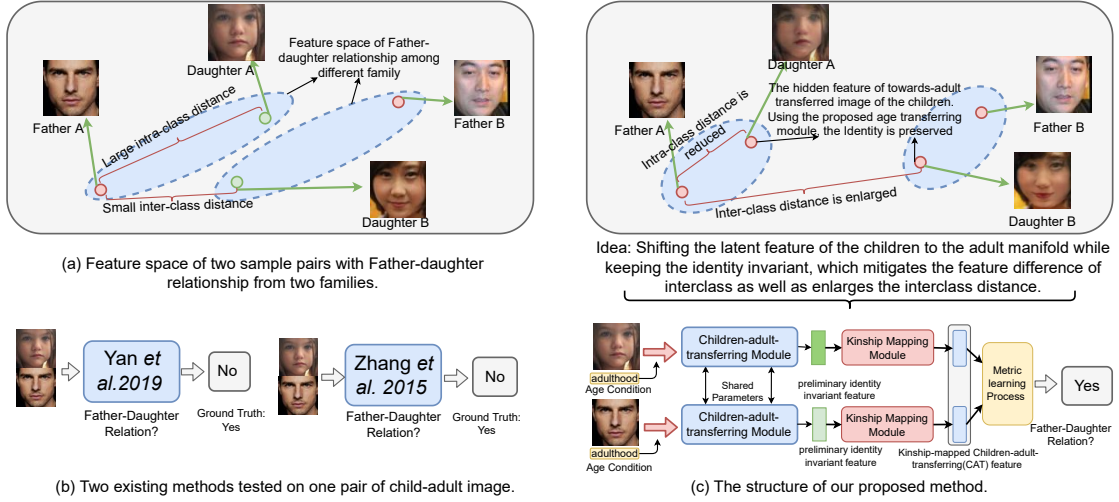


Figure 31: (a) Children may have larger appearance differences compared to their parents than to their peers of similar age. Consequently, the intra-class distance can be larger than the inter-class distance. (b) and (c) show the verification results of three methods (Yan et al. [232], Zhang et al. [240] and our proposed method) which are trained on the Nemo-Kinship-Children dataset. An external image pair is tested. The result shows that our proposed method can generate the correct verification result. (c) The novelty of our approach is to transfer images of children from childhood to adulthood while keeping the identity invariant at feature level. This approach enlarges the inter-class distance and reduces the intra-class distance.

- We introduce a novel challenge of kinship verification based on child-adult pairs and collect a new benchmark dataset of child-adult pairs for this specific task.
- We propose a Children-Adult-Transferring (CAT) Module and Kinship Mapping Module (KMM) using an attention mechanism.
- A new Children-Adult-Transferring Network (CATNet) is proposed.
- Large scale experiments are conducted. The experimental results show that the towards-adult transferred features of children images improve the representation of kinship relations and subsequently the kinship verification performance.

## 4.2 RELATED WORK

### 4.2.1 Kinship Verification

The first method of image based kinship verification is proposed by Fang *et al.* [62] in 2010. Then, Lu *et al.* propose a number of metric learning methods [128, 229] by minimizing the intra-class and maximizing the inter-class samples. Zhou *et al.* [251] use Gabor-based gradient orientation pyramid features to conduct kinship verification in uncontrolled circumstances. Other methods [112, 114, 199] focus on the use of neural networks. Zhang *et al.* use convolutional neural networks [240] demonstrating the effectiveness of exploiting deep learning with a limited set of samples and Yan *et al.* [232] add attention mechanisms in deep learning networks for kinship verification. Dibeklioglu *et al.* [48] explore video-based facial representations by transforming the

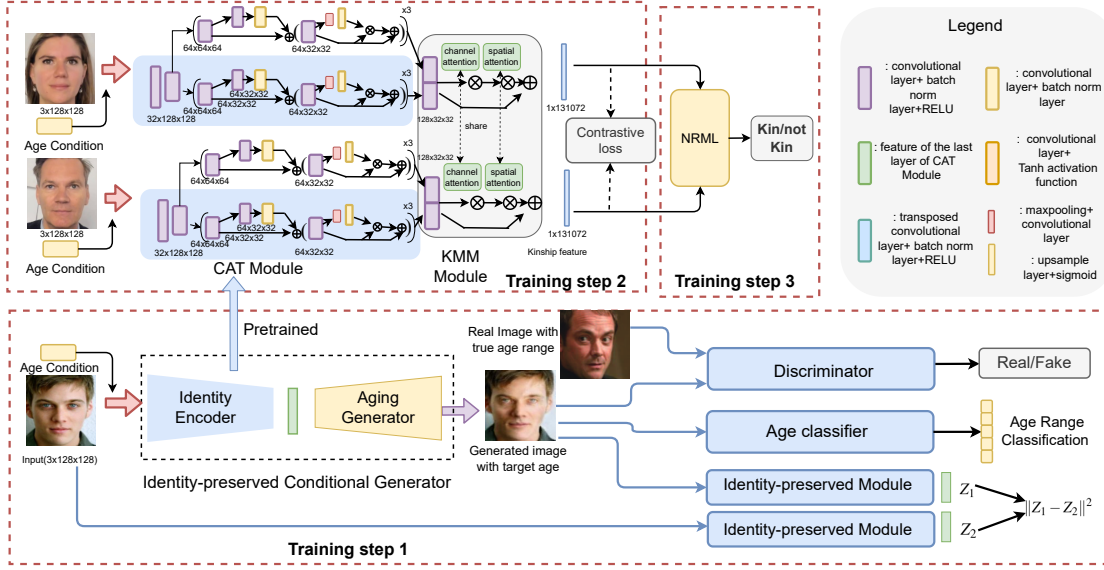


Figure 32: Pipeline of the proposed Children-Adult-Transferring Network (CATNet). The top left side of the figure shows an overview of the CAT Module. The boxes with dashed lines depict that the training process of the proposed approach is divided into three stages: the training of the Identity-Preserved Conditional Generator, the training of the Kinship Mapping Module and the training of the NRML.

facial appearance of kin-pairs using a deep contrastive learning architecture. As for age related work on kinship verification, Xia *et al.* [219] show that the aging process influences the kinship verification performance. Lelis *et al.* [108] conduct a large-age-variation-adapted method by using extracted features from a pre-trained model. Wang *et al.* [194] pre-trained a siamese model by using face attributes (*e.g.*, age, gender, kin-or-not) to predict the family tree in a photo. Recently, a towards-young cross-generation model is proposed by Wang *et al.* [197] by generating younger parent images from older ones. In contrast to [197], our approach mainly focuses on child-adult images and we exploit the hidden feature representation of an aging generator model instead of using the generated images.

#### 4.2.2 Age-Invariant Face Feature Learning and Cross-Age Face Synthesis

The aim of age-invariant feature learning is to reduce face variations but to keep the identity-related features. Li *et al.* use a local feature description [115] to compute the age-invariant information, and [206] use CNN's. Wang *et al.* [206] introduce the Orthogonal Embedding CNN (OE-CNN) model and decompose the age-related and identity-related information into two orthogonal components. Previous synthesis methods [162] require massive annotated data (*e.g.* facial shape, structure, and muscles) and are computationally expensive. Wang *et al.* [200] propose a face aging (RFA) framework based on a recurrent neural network. Song *et al.* [181] propose dual conditional GANs (Dual cGANs) for face aging. They use age estimation techniques to preserve the identity information. To make use of both recognition and synthesis, Zhao *et al.* [245] propose an Age-Invariant Model (AIM) leveraging both cross-age face synthesis and recognition. Similarly, Wang *et al.* [208] propose an Identity-Preserved Conditional Generative Adversarial Networks

(IPCGANs) framework and utilize pretrained AlexNet to preserve the identity information for the synthesized faces.

### 4.3 METHOD

#### 4.3.1 Problem Formulation and Motivation

Generally, kinship verification is a binary task aiming at verifying whether a pair of persons is kin or not. One of the difficulties of computing child-adult pairs is the difference in face outlines between children and adults. Another problem is the variation in skin color and texture between children and adults. With age, the texture and color of the face change. Such aging issues not only reduce the resemblance of child-adult kinship members but also influence the verification of pairs with large age influences.

To this end, we propose to exploit an age-transferring generator to learn the age-transferring information and at the same time to preserve the identity information. We approach the kinship verification problem from an age-transferring generative perspective by introducing towards-adult-kinship-related features.

#### 4.3.2 Pipeline

As shown in Figure 32, a novel Children-Adult-Transferring Network (CATNet) is proposed based on the Children-Adult-Transferring (CAT) module by using hidden features of the generator. The framework consists of two branches. The shared-weight CAT module computes the preliminary features separately from two test images with age range conditions. Then, the preliminary features are mapped onto a kinship related manifold by the Kinship-Mapping Module (KMM), which enhances the similarity features through an attention mechanism both element-wise and channel-wise. The mapped kinship features are further processed by the NRML.

#### 4.3.3 Identity-preserved Aging Generator

To simulate the facial aging process at feature space, we propose a novel Identity-Preserved Aging Generator as the basic generative model. To this end, the aim is to integrate the Identity-Preserved Aging Generator into an Identity Encoder (the bottom branch of Children-Adult-Transferring (CAT) Module) and an Aging Generator. Different from the existing IPCGANs [208], we focus on exploiting identity-preserved hidden features instead of creating aging figures.

#### *Sub-modules*

The aim of Children-Adult-Transferring (CAT) Module is to compute identity-related hidden features conditioned on the target age range. As shown in Figure 32, the CAT Module consists of two sub-branches. The top sub-branch consists of three convolutional layers and an attention block. The bottom sub-branch is the Identity Encoder, which is incorporated into the Identity-Preserved Aging Generator.

In our proposed architecture, test image  $x$  is fed into the CAT Module and then mapped onto the manifold of real image  $y$  within the target age range  $C_t$  by the Aging Generator  $G$ . In order to train the CAT Module in an adversarial manner, discriminator  $D$  is formed capturing the distribution of true images. The distributions of input  $x$  and target images  $y$  are denoted by  $p_x(x)$  and  $p_y(y)$ .

A modified LSGANs [135] loss is used during the training process:

$$\begin{aligned}\mathcal{L}_D &= \frac{1}{2} \mathbb{E}_{y \sim p_y(y)} \left[ (D(y|C_t) - 1)^2 \right] + \frac{1}{4} \mathbb{E}_{y \sim p_y(y)} \left[ (D(y|C_n))^2 \right] \\ &\quad + \frac{1}{4} \mathbb{E}_{x \sim p_x(x)} \left[ (D(G(E(x|C_t))|C_t))^2 \right], \\ \mathcal{L}_G &= \frac{1}{2} \mathbb{E}_{x \sim p_x(x)} \left[ (D(G(E(x|C_t))|C_t) - 1)^2 \right],\end{aligned}\tag{19}$$

where  $E, G$  and  $D$  indicate the Identity Encoder, Aging Generator and Discriminator respectively. The value  $C_t$  represents the following five age ranges: 11-20, 21-30, 31-40, 41-50, and 50+ respectively.

**IDENTITY-PRESERVED MODULE** To keep the person-dependent properties consistent, we use Alexnet network  $h(\cdot)$  pretrained on ImageNet as the Identity-Preserved Module to compute the latent feature space of test  $x$  and target image  $y$ . Then, the identity loss is defined by:

$$\mathcal{L}_{\text{identity}} = \sum_{x \in p_x(x)} \|h(x) - h(G(E(x|C_t)))\|^2.\tag{20}$$

**AGE DOMAIN CLASSIFIER** To support  $G(E(\cdot))$  to generate photo-realistic images in the target age domain, a pretrained age classifier is used by adapting the pretrained Alexnet on the CACD dataset [25]. The age loss  $\mathcal{L}_{age}$  is given by:

$$\mathcal{L}_{age} = \sum_{x \in p_x(x)} \ell(G(E(x|C_t)), C_t),\tag{21}$$

where  $\ell$  denotes the cross entropy loss of the pretrained age classifier.

#### *Objective Function of Identity-preserved Aging Generator*

As shown in Figure 32, an overview of the training scheme of the Identity-Preserved Aging Generator is given. Objective functions of  $D$  and  $G$  are defined by:

$$\begin{aligned}G_{\text{loss}} &= \lambda_1 \mathcal{L}_G + \lambda_2 \mathcal{L}_{\text{identity}} + \lambda_3 \mathcal{L}_{age}, \\ D_{\text{loss}} &= \mathcal{L}_D,\end{aligned}\tag{22}$$

where  $\lambda_1, \lambda_2$  and  $\lambda_3$  are the weights of each loss. In our framework, the Identity-Preserved Aging Generator is pretrained. The pretrained CAT Module is used as a preliminary feature extractor for the kinship verification task.

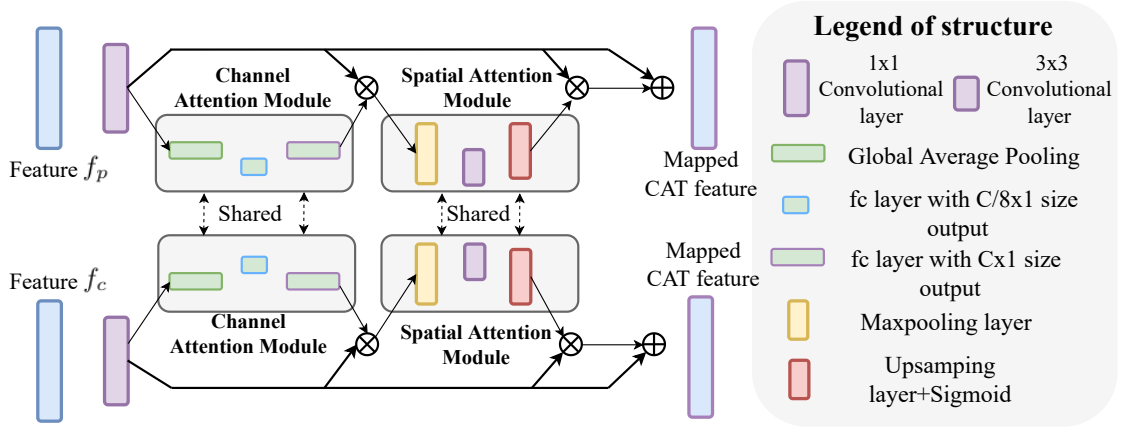


Figure 33: Detailed structure of the Kinship Mapping Module. Features  $f_p$  and  $f_c$  are mapped and enhanced by channel- and element-wise attention.

#### 4.3.4 Identity-Invariance-Aging-Transferring Network

##### Kinship Mapping Module

To augment the kinship-related information from the preliminary features, we add a Kinship Mapping Module (KMM) after the final layer of the CAT Module. As shown in Figure 33, the KMM aims to preserve the kinship related information and to make the kinship features of parents and children more distinguishable. These are extracted from the identity invariant information of the CAT Module containing spatial information and large channel wise features. To further enhance the kinship-related features, we add the attention module both channel-wisely and spatial-wisely.

Firstly, features  $f_p$  and  $f_c$  for channel  $C$  are calculated and mapped by two  $1 \times 1$  convolutional layers respectively. Then, a channel wise attention module is utilized based on the Squeeze-and-Excitation block [87, 149]. To reduce parameter overhead [213], the channel-wise attention map is squeezed into  $\mathbb{R}^{C/8 \times 1 \times 1}$  and excited back into  $\mathbb{R}^{C \times 1 \times 1}$ . Then, the aggregated channel attention is used by the spatial attention module to enhance the local spatial information. A model similar to [232] is utilized using a bottom-up top-down structure consisting of a Maxpooling layer,  $3 \times 3$  convolutional layers and one up sampling layer with Sigmoid function. Finally, the mapped Identity-Invariance-Aging features are obtained.

Further, the contrastive loss is used. Assuming the preliminary feature to be mapped as  $\mathbf{I}_p$  and  $\mathbf{I}_c$ , representing the preliminary features of a parent image and a child image respectively, the contrastive loss [80] is defined by:

$$\mathcal{L}(W, Y, \mathbf{I}_p, \mathbf{I}_c) = (Y) \frac{1}{2} (D_W)^2 + (1 - Y) \frac{1}{2} \{\max(0, m - D_W)\}^2, \quad (23)$$

where  $D_W$  is the parameterized distance function represented by the Euclidean distance between points on the mapped manifold:  $D_W(I_p, I_c) = \|G_{w_p}(I_p) - G_{w_c}(I_c)\|$ .  $G_{w_p}$  and  $G_{w_c}$  are the convolution layers of the kinship mapping module with parameters obtained by the training process.  $Y$  is the label for kinship relation.  $Y = 0$  indicates that  $\mathbf{I}_p$  and  $\mathbf{I}_c$  are dissimilar.



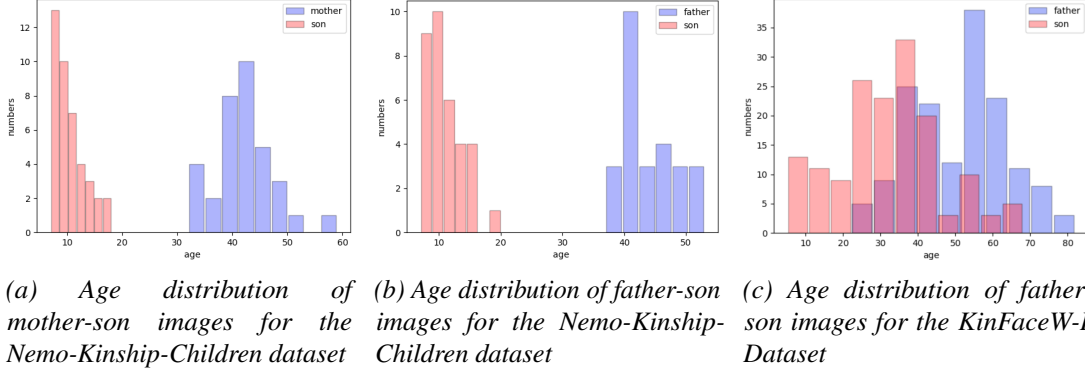


Figure 34: Age distribution of the Nemo-Kinship-Children Dataset. (c) The Age distribution of KinFaceW-I dataset has been annotated by two persons since there is no age label for the KinFaceW-I datasets.

#### Neighborhood Repulsed Metric Learning (NRML)

After training the KMM, kinship-preserved features  $\mathbf{p}_i$  and  $\mathbf{c}_j$  are computed. We use NRML [130] to further learn the distance metric. NRML is given by the following optimization strategy:

$$\begin{aligned} \max_A J(A) &= J_1(A) + J_2(A) - J_3(A) \\ &= \frac{1}{Nk} \sum_{i=1}^N \sum_{t_1=1}^k (\mathbf{p}_i - \mathbf{c}_{it_1})^T A (\mathbf{p}_i - \mathbf{c}_{it_1}) \\ &\quad + \frac{1}{Nk} \sum_{i=1}^N \sum_{t_2=1}^k (\mathbf{p}_{it_2} - \mathbf{c}_i)^T A (\mathbf{p}_{it_2} - \mathbf{c}_i) \\ &\quad - \frac{1}{N} \sum_{i=1}^N (\mathbf{p}_i - \mathbf{c}_i)^T A (\mathbf{p}_i - \mathbf{c}_i), \end{aligned} \quad (24)$$

where  $\mathbf{c}_{it_1}$  indicates the  $t_1$ -th  $k$ -nearest neighbor of  $\mathbf{c}_i$ , and  $\mathbf{p}_{it_2}$  indicates the  $t_2$ -th  $k$ -nearest neighbor of  $\mathbf{p}_i$ .  $\mathbf{c}_i$  and  $\mathbf{p}_j$  ( $i = j$ ) are taken as kinship related and  $\mathbf{c}_i$  and  $\mathbf{p}_j$  ( $i \neq j$ ) are non-kinship related.  $N$  is the number of training samples and  $A$  is a square matrix to be learned.

#### 4.3.5 Nemo-Kinship-Children Dataset

Table 17: Scales of different kinship datasets. Our newly collected Nemo-Kinship-Children dataset is comparable to other kinship related datasets both on kinship types, numbers of images, and pairs.

Dataset	F-D	F-S	M-D	M-S	pairs	people	type
CornellKin [62]	33	60	39	18	150	300	Image
UB Kinface ver1 [219]	-	-	-	-	270	180	Image
UB Kinface ver2 [221]	-	-	-	-	400	400	Image
KinFaceW-I [129]	134	156	127	116	533	-	Image
KinFaceW-II [129]	250	250	250	250	1000	-	Image
IIITD [98]	33	52	26	52	163	-	Image
<b>Nemo-Kinship-Children</b>	51	60	84	75	270	209	<b>Video</b>

Currently, many public kinship datasets contain unbalanced age images. Due to the small portion of children images, these datasets are not directly suited to assess the child-adult-related kinship verification task. To obtain child-related images, we created

Table 18: Performance of the different methods by 5-fold cross validation on the Nemo-Kinship-Children Dataset.

Methods	<i>F-D</i> ↑	<i>F-S</i> ↑	<i>M-D</i> ↑	<i>M-S</i> ↑	average accuracy ↑
Zhang <i>et al.</i> 2015 [240]	0.520	0.467	0.454	0.479	0.521
Yan <i>et al.</i> 2019 [232]	0.513	0.593	0.522	0.563	0.547
Lu <i>et al.</i> 2013 [131]	0.632	0.640	0.664	0.646	0.646
CATNet (ours)	0.718	0.733	0.652	0.779	0.721

the Nemo-Kinship-Children dataset. This dataset is a subset of the data collected in the Nemo-museum as part of Science Live, the innovative research program organized by Science Center NEMO<sup>1</sup>.

The Nemo-Kinship-Children Dataset comprises 3553 videos of 209 people. All images in the dataset are recorded indoors with consistent lighting conditions, and all subjects have frontal poses. The dataset can be divided into four groups: Father-Daughter (F-D), Father-Son (F-S), Mother-Daughter (M-D), and Mother-Son (M-S). The ages of the subjects vary from 7 to 63. The histogram in Figure 34 shows that the Nemo-kinship-Children dataset has a large number of children images under 16 compared to the public KinFaceW-I dataset. The portion of children images under 16 (without 16) are 92.8% (F-D), 97.1% (F-S), 95.7% (M-D), 97.6% (M-S), which results in a large variation in age. The table 17 shows the comparison of Nemo-Kinship-Children dataset with public datasets.

#### 4.4 EXPERIMENTS

##### 4.4.1 Data Selection and Preparation

Large scale experiments are conducted using the Nemo-Kinship-Children and KinFaceW-I datasets [129]. KinFaceW-I and KinFaceW-II are commonly used datasets for kinship verification. In KinFaceW-I, all kinship pairs are taken from different pictures. Since KinFaceW-II pairs are cropped from the same photo, the image pairs contain the same imaging conditions. Therefore, the KinFaceW-I dataset is selected. For Nemo-Kinship-Children Dataset, we extract one image per subject to make it comparable to other public datasets. The face of the subject is cropped, aligned, and resized to  $160 \times 160$ .

##### 4.4.2 Experimental Setups

The training of our proposed method is divided into three stages: the pre-training of the Identity-Invariance-Aging Module, the training of the KMM, and Neighborhood Repulsed Metric Learning. The training of the Identity-preserved Aging Generator follows the experiments in [208]. A Cross-Age-Celebrity Dataset (CACD) [25] is used for training. For the training of the KMM, the Nemo-Kinship-Children dataset is used. The scheme is illustrated in Figure 32. The trained Identity Encoder is frozen to prevent

<sup>1</sup> <https://www.nemosciencemuseum.nl/nl/wat-is-er-te-doen/activiteiten/science-live/>

Table 19: Ablation study: The accuracy results of different features with NRML on Nemo-Kinship-Children dataset. CAT+NRML (w/o aging) is using generated feature without aging children from childhood to adulthood.

Methods	<i>F-D</i> ↑	<i>F-S</i> ↑	<i>M-D</i> ↑	<i>M-S</i> ↑	average accuracy ↑
LBP+NRML	0.680	0.623	0.653	0.585	0.635
HOG+NRML	0.657	0.590	0.621	0.682	0.615
(HOG+LBP)+(m)NRML	0.632	0.640	0.664	0.646	0.646
CAT+NRML (w/o KMM)	0.685	0.673	0.642	0.742	0.686
CAT+NRML (w/o aging)	0.678	0.690	0.652	0.754	0.694
CAT+NRML (ours)	0.718	0.733	0.652	0.779	0.721

Table 20: Ablation study: The AUC results of different features with NRML on the Nemo-Kinship-Children dataset. CAT+NRML (w/o aging) is using generated features without aging children from childhood to adulthood.

Methods	<i>F-D</i> ↑	<i>F-S</i> ↑	<i>M-D</i> ↑	<i>M-S</i> ↑	average AUC ↑
LBP+NRML	0.545	0.470	0.534	0.622	0.543
HOG+NRML	0.518	0.562	0.563	0.505	0.537
(HOG+LBP)+(m)NRML	0.520	0.504	0.544	0.569	0.535
CAT+NRML (w/o KMM)	0.546	0.567	0.561	0.602	0.569
CAT+NRML (w/o aging)	0.566	0.613	0.597	0.623	0.600
CAT+NRML (ours)	0.584	0.623	0.596	0.634	0.609

from being disturbed by parameter updating of the Kinship Mapping Module. As mentioned above, a contrastive loss is used and  $m$  is set to be 10. Adam is adopted with a learning rate of  $10^{-4}$  and weight decay  $5 \times 10^{-3}$ . The batch size is set to be 36. The  $\lambda_{1-3}$  are set to be 70 : 1 : 1. After training the KMM, the obtained kinship-related features are used for NRML for further processing. The training of NRML follows the training procedure from the public source of [131].

#### 4.4.3 Comparison with Current Methods

To compare our approach, we select three representative methods: Part-aware attention network [232], CNN-point Network [240], and (m)NRML [131] with LBP and HOG features. (m)NRML is a typical kinship verification method in metric learning. It is often used and the source code is publicly available. We use the (m)NRML method with LBP and HOG features. According to the provided code of NRML, cosine similarity is used for evaluating the similarity of test samples. The accuracy of each kinship type is the average result of the best performance on each validation fold. Zhang *et al.* [240] representative for deep learning methods for kinship verification. Part-aware attention network is a recently proposed method using ensembles of CNNs with an attention module. These two methods show competitive performance compared to previous handcrafted-feature based methods. Since public source code is not available, we have re-implemented the fundamental architecture of the part-aware attention networks (attention-only network [232]) and the fundamental architecture (CNN-Basic Network) of Zhang *et al.*'s method from scratch following [232, 240].

### Qualitative Results

Table 21: Accuracy of different methods for 5-fold cross validation on the KinFaceW-I dataset.

Methods	<i>F-D</i> ↑	<i>F-S</i> ↑	<i>M-D</i> ↑	<i>M-S</i> ↑	average accuracy ↑
Yan <i>et al.</i> 2014 Discriminative [228]	0.675	0.705	0.720	0.655	0.689
Yan <i>et al.</i> 2014 Prototype [230]	0.735	0.675	0.661	0.731	0.701
Dehghan 2014 Who [44]	0.764	0.725	0.719	0.773	0.745
Zhang <i>et al.</i> 2015 [240]	0.709	0.770	0.795	0.689	0.741
Yan <i>et al.</i> 2019 [232]	0.616	0.725	0.772	0.608	0.680
Lu <i>et al.</i> 2013 (LBP) [131]	0.702	0.805	0.741	0.685	0.733
Lu <i>et al.</i> 2013 (HOG) [131]	0.709	0.782	0.733	0.698	0.731
Lu <i>et al.</i> 2013 (LBP+hog) [131]	0.717	0.802	0.740	0.715	0.743
Liang <i>et al.</i> 2019 Weighted [117]	0.785	0.739	0.806	0.819	0.787
Yan 2019 Learning [225]	0.796	0.736	0.761	0.815	0.776
Li <i>et. al.</i> 2020 Graph [114]	0.732	0.795	0.862	0.780	0.792
Wang <i>et.al.</i> 2020 Discriminative [199]	0.765	0.77	0.852	0.758	0.786
Yan <i>et al.</i> 2021 Multi [231]	0.850	0.875	0.881	0.809	0.856
Li <i>et.al</i> 2021 Reasoning [112]	0.788	0.817	0.814	0.886	0.826
Chen <i>et.al</i> 2022 Deep [31]	0.837	0.800	0.822	0.864	0.831
CATNet (ours)	0.780	0.840	0.804	0.780	0.801

Table 22: Ablation study: The accuracy results of different features with NRML on the KinFaceW-I dataset.

Methods	<i>F-D</i> ↑	<i>F-S</i> ↑	<i>M-D</i> ↑	<i>M-S</i> ↑	average accuracy ↑
LBP+NRML	0.7017	0.8046	0.7401	0.6851	0.7329
HOG+NRML	0.7091	0.7821	0.7333	0.6982	0.7307
(HOG+LBP)+(m)NRML	0.7165	0.8015	0.7399	0.7154	0.7433
CAT+NRML (w/o KMM)	0.7165	0.7725	0.7801	0.7281	0.7493
CAT+NRML (w/o aging)	0.7387	0.8015	0.7487	0.7585	0.7619
CAT+NRML (ours)	0.7799	0.8398	0.8039	0.7803	0.8010

### Quantitative Comparison

The results of the 5-fold cross validation on the Nemo-Kinship-Children dataset are shown in Table 18. Our approach outperforms NRML and the two deep learning methods. The results also show that the methods by Yan *et al.*'s and Zhang *et al.*'s yield relatively low accuracy on the Nemo-Kinship-Children dataset. We attribute this to the size of the Nemo-Kinship-Children dataset as well as the large discrepancy among children and adults.

The generated images from different datasets are shown in Figure 35. The results show apparent aging changes between the original child's image and the transferred one of our newly collected dataset, and the FIW and CornellKin dataset. The texture of the transferred image face is more similar to the adult texture.

Table 23: Ablation study: The AUC results of different features with NRML on the KinFaceW-I dataset.

Methods	$F-D \uparrow$	$F-S \uparrow$	$M-D \uparrow$	$M-S \uparrow$	average AUC $\uparrow$
LBP+NRML	0.7017	0.8325	0.7290	0.6861	0.7373
HOG+NRML	0.7344	0.8589	0.7658	0.6859	0.7612
(HOG+LBP)+(m)NRML	0.7288	0.8582	0.7565	0.6957	0.7598
CAT+NRML (w/o KMM)	0.7268	0.7759	0.7844	0.771	0.7645
CAT+NRML (w/o aging)	0.7268	0.7759	0.7844	0.771	0.7645
CATNet (new)	0.7797	0.8275	0.8356	0.8164	0.8148

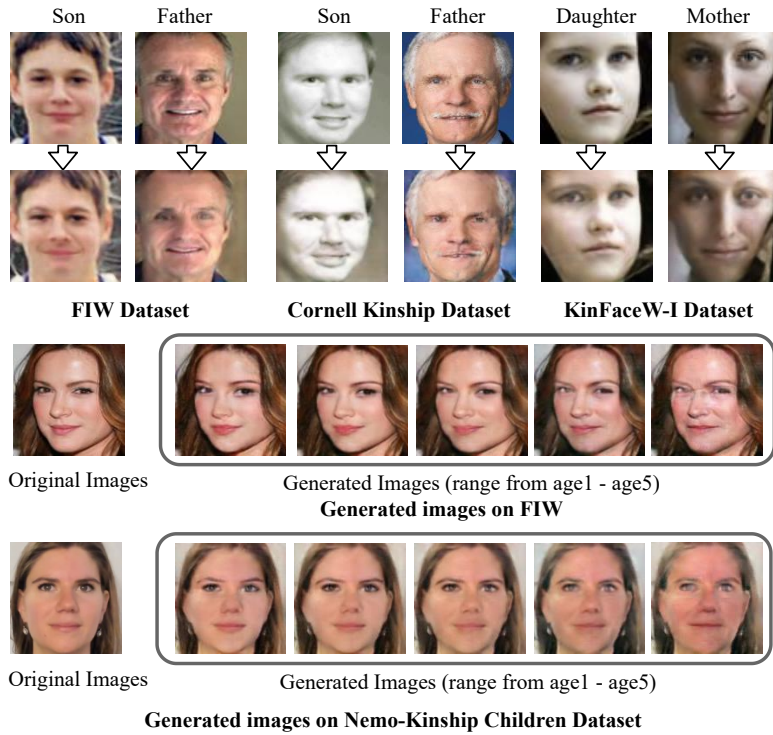


Figure 35: Qualitative results of generated images on a target age range while preserving identity.

#### 4.4.4 Ablation Study

We compare the CAT features of our approach with the HOG and LBP features. The results show that the NRML with CAT features outperform HOG and LBP. For the Kinship Mapping Module, the results of Table 19 show that the Kinship Mapping Module extracts kinship related features improving the final results in both Nemo-Kinship-Children dataset. We also compare the aged and non-aged features used by our proposed architecture. It shows that the transferred hidden feature improves the kinship verification performance. The AUC results in Table 20 show that our approach, based on CAT features, produces the highest average AUC.

#### 4.4.5 Robustness and Generalization

In this section, we evaluate our approach to show the robustness and generalization of our model. The results of the different methods on KinFaceW-I is shown in Table 21. It is shown that our approach outperforms the two re-implemented methods and NRML with LBP and HOG features and obtains more robust results than the methods by Yan *et al.*'s and Zhang *et al.*'s [232, 240]. In comparison to current state-of-the-art methods, our method is compatible to Li *et al.* [114] and Wang *et al.* [199]. The difference in performance with respect to Yan *et al.* [231] is that our method is focused on young children-related pairs, while the KinFaceW-I also contains images of different ages. Table 22 and Table 23 are consistent with the previous experimental results. The results in Table 22 show the performance of the Kinship Mapping Module. The AUC results in Table 23 show the suitability of our CAT feature.

#### 4.5 CONCLUSION

In this chapter, we proposed a novel Identity-Invariance-Aging-Transferring approach based on newly designed modules. A kinship mapping module is used to compute the improved kinship-related information from the features of the CAT Module. The results show that, compared to the handcrafted feature, the transferred features capture the hidden features of genetic relationships and provide more robust results for child-related and elderly-related pairs.

---

## KINSHIP SIMILARITY FOR OPEN SETS

---

### 5.1 INTRODUCTION

Image-based kinship recognition [165] aims at determining the genetic relationship between people by analyzing images of their faces. In recent years, more and more attention has been focused on kinship related tasks such as kinship verification, family recognition, and kinship identification [166] due to many applications such as searching missing children [217], family album organization, forensic investigation [170], crime scene investigation [100], social media [236] and behavior analysis [240], historical and genealogical research [95], and automatic image annotation [130].

The recognition of kinship relationships between people as a hierarchical (tree) structure is an important task because people may share origins at different degrees. The genetic (kinship) relationship between two people is the amount of DNA they have in common because they are related. According to the average percent of DNA shared between relatives<sup>1</sup>, some kinship relations may have similar gene overlap. As illustrated in Figure 36, family members have various degrees of kinship relationships. For example, the boy has 50% genetic origin with his (full) father and mother. In addition, the genetic relationship between the boy and his (full) brother is 0.5, as they share an average of 50% origins of their DNA. The boy and his grandmother share 25% DNA. Hence, genetic relationships between family members are determined by a family tree. Obviously, outside the family, there exists a large set of people who have no genetic origin at all (or very far) with the family members. Hence, kinship recognition is by definition an open set problem for real-life scenarios. However, the recognition of different degrees of kinship in open set scenarios has largely been ignored so far.

Therefore, this chapter focuses on kinship at various degrees for open collections (*i.e.* including kin and non-kin related people). The aim is to determine family relationships and their corresponding degrees of kinship hierarchically. To this end, we propose a novel and more general task called the Open-set Kinship Similarity Measurement (OKSM). As illustrated in Figure 36, similarities between images of faces of people are derived to measure different degrees of kinship relationship *i.e.* the first-degree relationship between parent and child, the second degree between grandparent and grandchild, *etc.* The more genetically related people are, the higher the similarity. Furthermore, current open set classification methods are mainly focused on distinguishable classes (not pairs), and most

---

<sup>1</sup> [https://isogg.org/wiki/Autosomal\\_DNA\\_statistics](https://isogg.org/wiki/Autosomal_DNA_statistics)

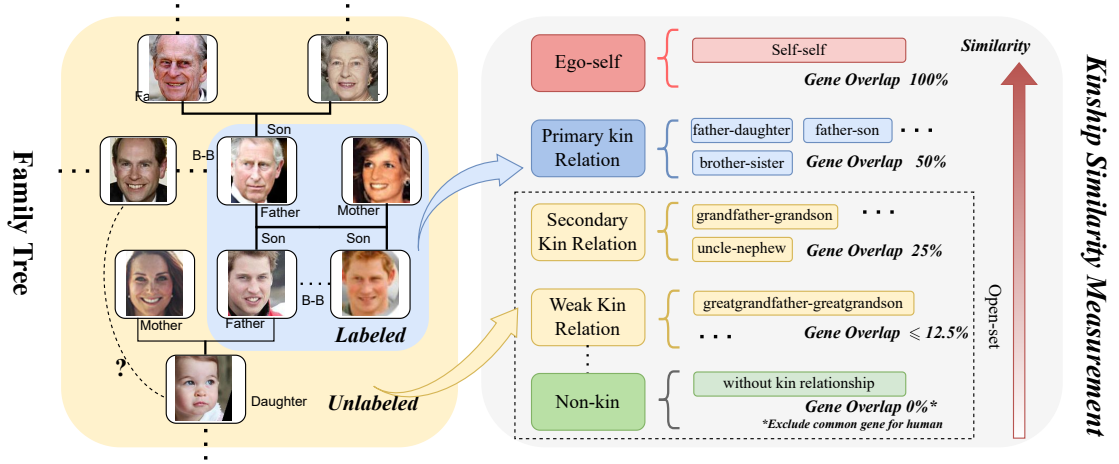


Figure 36: The proposed open set kinship recognition task. The open set contains both (1) family members (having various degrees of kinship) and (2) non-kin persons. Kinship recognition is by definition an open set problem for real-life scenarios.

of the benchmark datasets are image-based [69]. In contrast, our method is pairwise-based and exploits mutual information from positive pairs by merely re-matching them. Moreover, we propose a data-driven strategy to compute hierarchical kinship degrees. Firstly, our method assigns images (Primary-Kin) into self-self (ego-self), kin-related (containing 50% genetics), and known-negative (unrelated) pairs. Then, a hierarchical kinship triplet loss is proposed to learn the similarity of the different pairs. Image pair features are projected onto a distance-based feature space. The distance measure is used to differentiate between primary kin, secondary kin, and non-kin relationships.

The main contributions of this chapter are summarized as follows:

1. Kinship similarity is introduced for open sets.
2. A method is proposed to derive pairwise information by re-matching known (positive) pairs.
3. A hierarchical kinship network is proposed to distinguish primary kin, secondary kin, and non-kin relationships for open-set kinship scenarios.
4. The proposed method outperforms state-of-the-art methods on the FIW dataset.

## 5.2 RELATED WORK

### 5.2.1 Kinship Recognition and Related Tasks

Image-based kinship recognition receives more and more attention in the computer vision community [62, 165, 195]. Kinship recognition typically includes kinship verification, kinship identification, and family recognition [166]. Kinship verification is a binary classification problem, determining whether two persons, represented by image pairs of their faces, are kin related or not [165]. Different methods are proposed in the field of kinship verification [31, 50, 78, 112–114, 130, 219, 226, 228, 231, 251]. Kinship verification has limited applicability because it is not able to distinguish different kin-types. In



contrast, kinship identification aims to determine the different kinship types [166]. For example, Guo *et al.* [78] propose a graph-based method recognizing the type of kinship using pairwise kinship information. Wang *et al.* [194] propose a deep kinship recognition framework to predict kin relationships using family trees. Different feature modalities are used such as kin-or-not, gender, and relative age attributes. Recently, Wang *et al.* propose a kinship identification method based on joint learning [202]. The method is specifically designed for pairwise kinship identification. However, the above kinship verification and identification methods are used for closed sets. Kinship recognition is an open set problem in real-life scenarios, as many people are not (directly) kin related. Therefore, in this chapter, we propose a new and more general open-set kinship similarity framework.

### 5.2.2 Open-set Recognition

Open-set recognition (OSR) is the problem of properly handling unknown samples during the classification of the known ones [26, 27, 123, 171]. Based on traditional machine learning methods, previous OSR methods utilize SVM [172], extreme value machines (EVM) [77], and sparse representation-based methods [237]. Lately, deep neural network-based methods are proposed. An OpenMax layer is proposed by Bendale and Boulton [14] to circumvent the problem of the Softmax cross-entropy loss [69]. The method incorporates the likelihood of the recognition failures and adopts the concept of Meta-Recognition [242]. In addition, generative models are proposed [68, 92]. For example, the G-OpenMax algorithm [68] uses synthesized (fake) unknown samples. Neal *et al.* [143] generate known-negative images as counterfactual images to improve the robustness of their model. However, the performance can be negatively affected by the (poor) quality of the generated (fake) images. Recently, the class anchor clustering (CAC) loss is proposed [137] by clustering known classes into predefined clusters. Unlike the OpenMax method, the CAC loss predefines anchors of known classes without updating them during training. More recently, Chen *et al.* [27] propose Reciprocal Point Learning (RPL) utilizing reciprocal points representing extra-classes for one specific class. The method is extended to Adversarial Reciprocal Point Learning (ARPL) by adding an instantiated adversarial enhancement process [26]. The method obtains state-of-the-art performance for different datasets [104]. None of the open-set methods is used for kinship recognition. Since our new task is an open-set problem, we propose a new pairwise open-set method for kinship similarity.

## 5.3 PROBLEM FORMULATION AND COMPARISON

### 5.3.1 Problem Formulation

We formulate the open-set kinship problem as a classification problem. Face images are pre-processed in a pairwise form  $[(I_A, I_B), y]$ , where  $I_A$  and  $I_B$  denote two different images with kinship label  $y$ . In this way, a dataset  $\{[(I_{A_i}, I_{B_i}), y_i] \mid i = 1, 2, \dots, n\}$  is obtained where  $i$  is the index to the paired sample and  $y$  indicates the type of kinship. Further, a feature embedding neural network  $\phi$  is used to transform image pairs to vector

pairs, *i.e.*,  $(\phi(I_A), \phi(I_B)) = (\mathbf{x}_A, \mathbf{x}_B)$ .  $\mathcal{D} = \{[(\mathbf{x}_{A_i}, \mathbf{x}_{B_i}), y_i] \mid i = 1, 2, \dots, n\}$  is the resulting dataset.

Based on kinship similarity, the samples are categorized into three classes  $\{P, S, N\}$ , where  $P$ ,  $S$  and  $N$  represent the Primary-kin ( $P$ ), Secondary-kin ( $S$ ), and Non-kin ( $N$ ) relationships respectively. These three classes contain different kinship types with label values  $Y = \{\underbrace{1, 2, \dots, 7}_P, \underbrace{8, \dots, 11}_S, \underbrace{12}_N\}$ . Specifically, 1, 2, ..., 12 in  $Y$  represent

Father-Daughter (F-D), Father-Son (F-S), Mother-Daughter (M-D), Mother-Son (M-S), Brother-Brother (B-B), Brother-Sister (B-S), and Sister-Sister (S-S), Grandfather-Granddaughter (GF-GD), Grandfather-Grandson (GF-GS), Grandmother-Granddaughter (GM-GD), Grandmother-

Grandson (GM-GS), and pairs without kinship respectively. Since most existing kinship datasets only contain 1 – 7 different kinship types, only samples with label  $P$  are used for training, and samples with labels  $P, S, N$  are used for testing. In our settings, a measurement model is used to compute the kinship similarity of three classes. Similarity distance  $d$  is used to distinguish  $P$ ,  $S$ , and  $N$ .

### 5.3.2 Comparison with Kinship Recognition

Our problem is to determine three categories  $P$ ,  $S$ , and  $N$  by computing kinship-related information from pairs. Different from existing kinship recognition, our approach yields kinship similarity types. The aim of existing kinship recognition is to bring positive pairs together and negative pairs apart. However, this strategy may have a negative influence on the computation of hierarchical kinship relationships as it may lead to incorrect predictions of similarity types of the test samples. Therefore, our approach focuses on measuring the hierarchical kin similarities of pairs depending on their genetic sharing.

The Softmax Cross-Entropy is used as the standard loss in deep learning-based kinship recognition or related kinship tasks. Denoting  $\mathbf{z}$  as the extracted logits from a neural network, then the Softmax Cross-Entropy loss is defined by:

$$L = -\frac{1}{N} \sum_i^N \log \left( \frac{e^{\mathbf{z}_i}}{\sum_j e^{\mathbf{z}_j}} \right), \quad (25)$$

where  $\mathbf{z}_j$  represents the  $j$ -th element of the logits, and  $\mathbf{z}_i$  is the target logit [153] of the ground truth.  $N$  is the number of training samples. However, the close-set Softmax is not able to properly handle unknown samples because:

1. The Softmax Cross Entropy loss is not injective [65]. It can not guarantee a proper clustering behaviour [137].
2. The Softmax Cross Entropy [69] inherently has a closed set nature and can easily be misled by unknown samples [14].

On the contrary, our approach is specifically designed for open-set scenarios.

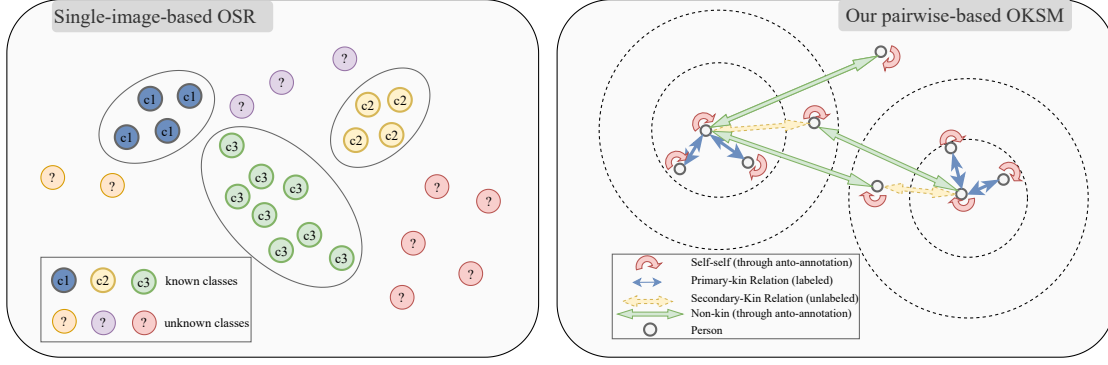


Figure 37: Left: General OSR methods [69] are single-based and known samples are independent of each other. Right: Our OKSM task is pairwise-based. Labels denote the relationships between image pairs.

### 5.3.3 Comparison with Open-set Recognition

Open-set methods are specifically designed to deal with unknown samples. However, current open set methods mostly focus on the classification of (single) classes [69]. As illustrated in Figure 37, the current open-set recognition approach mainly focuses on learning representative features for each known class. However, our open-set kinship approach is pairwise. The classification is based on the mutual relation between pairs. Hence, our model computes corresponding feature relationships among pairs. Moreover, our pairwise kinship model inherently obtains extra information by merely re-matching the known positive pairs. The right side of Figure 37 shows that after our data-driven strategy, the self-self and non-kin relationships can be auto-annotated and utilized for hierarchical feature representation. Standard open-set recognition methods consider two separate classes *i.e.* known  $K$  and unknown  $U$ . Let the open set be denoted by  $O_k = S^o - S_k$ ;  $S^o$  is the overall measure space [69], and  $S_k$  is the embedding space [26] of  $K$ , then  $O_k^U$  is the unknown space of  $U$ , the open space risk can be quantitatively described by:

$$\mathcal{R}_o(\psi_k, O_k^U) = \frac{\int_{O_k^U} \psi_k(x) dx}{\int_{S_k \cup O_k} \psi_k(x) dx}, \quad (26)$$

where  $\psi_k(x)$  is a binary measurable function. However, our open-set kinship problem is to separate  $P$ ,  $S$  and  $N$ . Hence, following the previous notion [26, 69], we re-define our open-set kinship measurement risk  $\mathcal{R}_{ko}$  as follows:

$$\mathcal{R}_{ko}(\psi_k, O_k^{U_N}) = \frac{\int_{O_k^U} \psi_k(x) dx}{\int_{S_k \cup O_k} \psi_k(x) dx} + \frac{\int_{O_k^{U_N}} \psi_{u_k}(x) dx}{\int_{S_k \cup O_k} \psi_{u_k}(x) dx}, \quad (27)$$

where  $\psi_{u_k}(x)$  is a binary measurable function,  $\psi_{u_k}(x) = 1$  indicates that  $U_N$  is regarded as  $U_k$ .  $U_N$  is the measure space for  $N$  samples and  $U_k$  is the measure space for  $S$  samples.

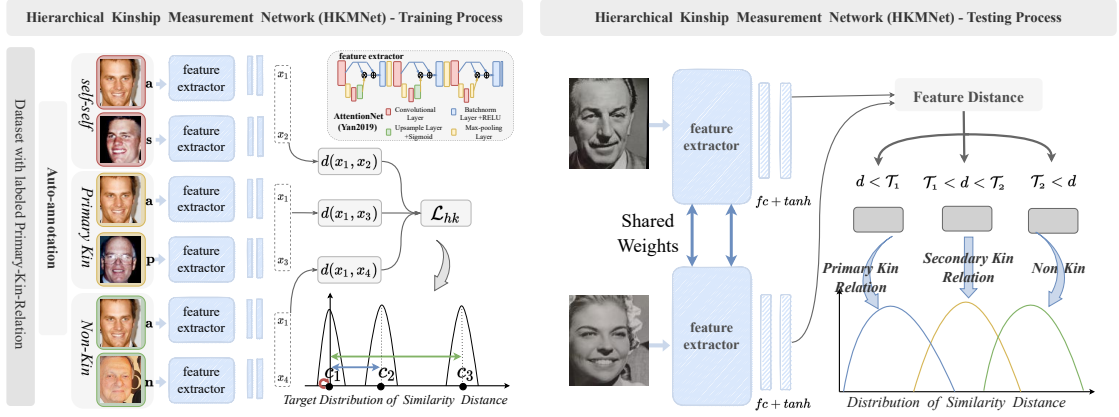


Figure 38: Our Hierarchical Kinship Measurement Network (HKMNet).

## 5.4 METHODOLOGY

The aim is to determine family relationships and their corresponding degrees of kinship hierarchically. To this end, we now introduce the architecture of the Hierarchical Kinship Measurement Network (HKMNet). The proposed hierarchical kinship network is a new 6-branch architecture integrating a modified  $fc$  layer,  $\tanh$  layers and similarity values  $\mathbf{c}$ . As illustrated in Figure 38, a data-driven strategy is adopted to extract hierarchical similarity from positive samples. Then, feature similarities of different hierarchical pairs are learned by our hierarchical kinship triplet loss. The final classification of test samples is dependent on the similarity measurements.

### 5.4.1 Hierarchical Information

Different from person re-identification or face retrieval, our new task predicts multiple classes. We only use Primary-Kin pairs as training samples. Secondary-Kin pairs are used as testing subsets for the open-set environment. Pairwise images inherently contain known-negative pairs of information. For example, as depicted in Figure 39, during training, known positive pairs are separated and re-matched into new pairs. Pairs containing two images of the same person form self-self pairs. Pairs containing known-positive images are denoted by known-positive (Primary-Kin relation) pairs, and pairs containing two images from two different families correspond to known-negative (Non-kin relation) pairs. Based on genetic similarity, self-self (ego-self) pairs correspond to the highest similarity (100% genetic overlap). Known-positive samples (Primary-Kin containing father-son, mother-son, father-daughter, mother-daughter) relates to medium similarity (around 50% genetic overlap), and known-negative (Non-kin) pairs correspond to the lowest similarity.

In this chapter, the auto-annotation process  $M(\cdot)$  is formulated as follows:

$$\{(\mathbf{a}, \mathbf{s}, \mathbf{p}, \mathbf{n}) | i = 1, 2, \dots, n\} = M(\{(I_{A_i}, I_{B_i}) | i = 1, 2, \dots, n\}), \quad (28)$$

where the  $i$ -th tuple  $(\mathbf{a}, \mathbf{s}, \mathbf{p}, \mathbf{n})_i$  represent the anchor image of the person, the image of the same person, the image of a person with a known positive class, and the image from

**Algorithm 1:** Pipeline of the Hierarchical Kinship Measurement Network.

---

**Input:** Image pairs  $(I_{A_i}, I_{B_i}) = \{(I_{A_i}, I_{B_i}) \mid i = 1, 2, \dots, n\}$

**Output:** *Primary-Kin, Secondary-Kin, Non-Kin*

**Training** (*only using Primary-Kin*)

```

while  $\mathcal{L}_{hk} > \Delta$  do
  batch images
   $= \{(\mathbf{a}, \mathbf{s}, \mathbf{p}, \mathbf{n})_i \mid i = 1, 2, \dots, N\}$ ,  $N =$ 
  batch size;
  if do hard example mining ( $\mathcal{H}_{\mathcal{E}}$ ) then
    Concatenate (batch images, hard
    images);
    calculate  $\mathcal{L}_{hk}$ ;
    select hard images with smaller
    values of  $d(x_1, x_4) - d(x_1, x_3)$ ;
  else
    calculate  $\mathcal{L}_{hk}$ ;
  end
end

```

**Testing**

```

if  $d(I_{A_i}, I_{B_i}) > \mathcal{T}_1$  then
  if  $d(I_{A_i}, I_{B_i}) < \mathcal{T}_2$  then
    Return Secondary-Kin ( $S$ )
  else
    Return Non-Kin ( $N$ )
  end
else
  Return Primary-Kin ( $P$ )
end

```

---

a different family, respectively. The total number  $n$  of re-generated sets is the same as the previous known positive sets.

#### 5.4.2 Distance-based Losses

After generating new sets from known-positive (Primary-Kin) pairs, the images are given to the feature extractor. The extracted features are then projected into a kinship-feature space. To learn hierarchical similarities instead of separating "positive" and "negative" samples, the Hierarchical Kinship Triplet Loss is used.

##### *Hierarchical Kinship Triplet Loss*

Since the same kinship category (Primary-Kin, Secondary-Kin, *etc.*) has the same gene overlap<sup>2</sup>, the kinship similarity is the same for these kinship categories. We first pre-set the similarity value  $\mathbf{c} = \{c_1, c_2, c_3\}$  for each training pair. Obviously, the feature distance of self-self pairs ( $c_1$ ) should be smaller than that of known-positive pairs ( $c_2$ ). And the feature distance of known-positive pairs ( $c_2$ ) should be smaller than the known-negative (Non-kin) pairs ( $c_3$ ). To learn such hierarchical information, the hierarchical kinship triplet loss is defined as follows:

$$\mathcal{L}_{hk} = \sum_{i=1}^3 \|d(x_1, x_{i+1}) - c_i\|^2 + \sum_{i=1}^2 \mathcal{L}_t(x_1, x_{i+1}, x_{i+2}), \quad (29)$$

<sup>2</sup> <https://customer.care.23andme.com/hc/en-us/articles/212170668-Average-Percent-DNA-Shared-Between-Relatives>

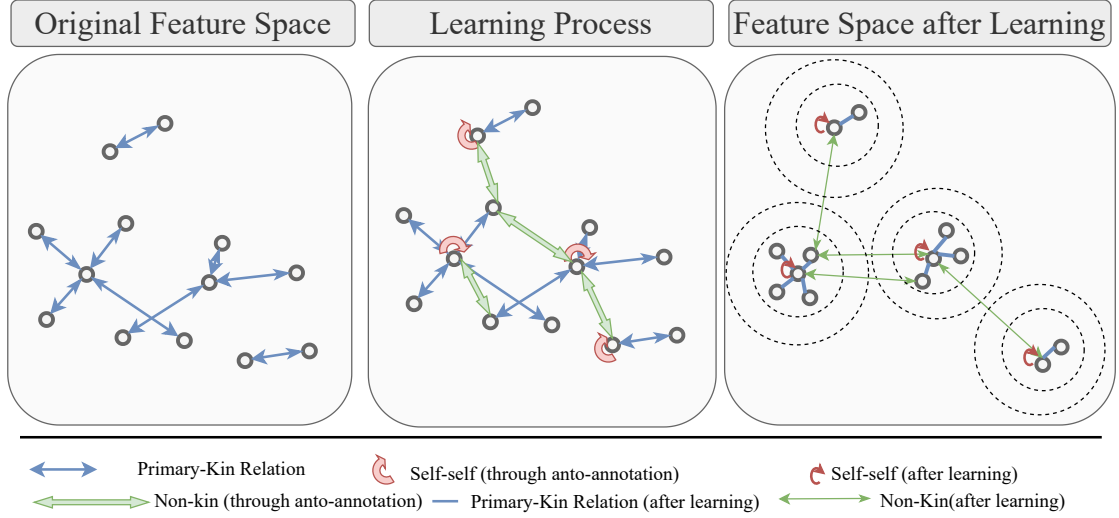


Figure 39: Feature learning pipeline of our proposed Hierarchical Kin-ship Measurement Network (HKMNet).

where the distance function is given by  $d(\cdot)$ , and  $x_1, x_2, x_3, x_4$  representing the features of **a**, **s**, **p**, **n** respectively.  $\mathcal{L}_t$  is the triplet loss represented by:

$$\mathcal{L}_t(x_a, x_p, x_n) = \max\{d(x_a, x_p) - d(x_a, x_n) + m, 0\}, \quad (30)$$

where  $m$  is a margin, and  $d(\cdot)$  is the Euclidean distance.  $x_a, x_p, x_n$  represent the anchor, positive, and negative feature of a specific person. Different from previous triplet pairs (anchor, pos, neg) [145], our triplet loss use the pairs with kin-based hierarchical information (self-self, known-positive, known-negative). The full pipeline is given in Algorithm 1.

## 5.5 EXPERIMENT

### 5.5.1 Datasets

The Families In the Wild (FIW) [166] dataset is used as the training and testing dataset. FIW is the largest publicly available image-based dataset for the kinship-related tasks. For our task, we downloaded the newest version (*rfiw2020*) from the official website. The images of FIW are collected from the Internet. The dataset contains 11193 unconstrained family photos of 1000 families. These images are processed into 418060 pairs with 11 kinship types. Since our task is new, the settings of training and testing datasets are different from previous settings. We follow the 5-split protocol and fit the dataset into our task<sup>3</sup>. We only labeled the Primary-Kin images. The training set contains one subset:

- Primary-Kin set, which only contains direct kinship types.

<sup>3</sup> We signed and strictly follow the ‘Social Safety Support Guide and Rules of Code of Conduct’ for using the human data.

Note that the Secondary-Kin images are not labeled and hence the problem becomes an open set problem. One of our tasks is to recall unlabeled data containing kinship information more effectively. Consequently, the Secondary-Kin set is not used for training and the training set corresponds to Primary-Kin label containing seven or four known positive kinship types: F-D, F-S, M-D, M-S, B-B, B-S, and S-S. These seven types are the mostly used kinship types in the literature. The testing set contains three subsets:

- Primary-Kin set, which contains known kinship types.
- Secondary-Kin (Unlabeled but kin-related) set, containing four kin type pairs (GF-GS, GF-GD, GM-GS, GM-GD) as unknown and unlabeled samples in real scenarios.
- Non-Kin set, which contains unknown pairs without kinship types.

More information (images and code) is anonymously available at our Github.

### 5.5.2 Experimental Settings

During data preprocessing, the images are resized to  $64 \times 64$ . Only horizontal flipping is applied for the purpose of data augmentation. Our proposed HKMNet is a plug-and-play network which is suited for different existing neural networks. In this section, AttentionNet [225] is used as the backbone. Experimental results with other feature extractors are provided in the supplementary material. The  $fc$  layer is shared-weighted and the  $\tanh$  activation function is added after the  $fc$  layer. The length of the features  $(x_1, x_2, x_3, x_4)$  are  $1 \times 64$ . Our HKMNet model is trained with SGD. The learning rate is set to be 0.001 and the epoch to be 40. During training, the hierarchical kinship triplet loss is calculated and hard examples ( $\mathcal{H}_E$ ) are reused. Margin  $m$  in  $\mathcal{L}_t(x_1, x_2, x_3)$  is set to be 0.5. Further,  $m$  in  $\mathcal{L}_t(x_1, x_3, x_4)$  is set to be 3. The similarity values  $c_1, c_2, c_3$  are set to be 0, 1, and 4 respectively.

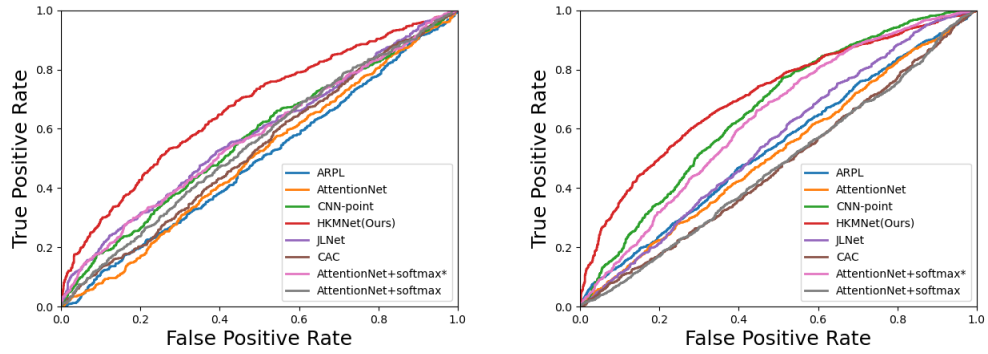
### Comparison

Unfortunately, publicly available codes are very limited in kinship recognition. Note that none of the existing methods can directly be applied to our new task. Therefore, all the compared methods are adapted to the new task at hand by changing their outputs accordingly.

To evaluate the performance of our proposed method, we compare our method with state-of-the-art methods in two different domains: (1) open-set recognition and (2) kinship-recognition. The CAC method [137] is one of the best performing distance-based open-set recognition methods. It explicitly trains known classes to form clusters around anchored class-dependent centers in logit space. ARPL is the state-of-the-art approach of open-set recognition tasks for many datasets. It models the unexploited extra-class space with the concept of Reciprocal Point and uses an instantiated adversarial enhancement process.

Table 24: Kinship similarity performance of 5-cross-validation on the FIW dataset. Here 7 kin-types ( $F-D$ ,  $F-S$ ,  $M-D$ ,  $M-S$ ,  $B-B$ ,  $B-S$ ,  $S-S$ ) in Primary-Kin category are used for training. The Unlabeled includes both Secondary-Kin and Non-Kin.

Methods		Accuracy	AUROC Primary-Kin vs Unlabeled	AUROC Secondary-Kin vs Non-Kin	F1
Same-backbone	AttentionNet+ <i>Softmax</i> * [225]	0.3752	0.6110	0.5780	0.3488
	AttentionNet+ <i>Softmax</i> [225]	0.3500	0.4842	0.5156	0.2835
Open-set Recognition	CAC [137]	0.3084	0.4893	0.5195	0.2264
	ARPL [26, 27]	0.3636	0.5388	0.4717	0.2590
kinship Related methods	CNN-point [240]	0.4266	0.6413	0.5865	0.3832
	AttentionNet* [225]	0.2926	0.5182	0.4802	0.2242
	JLNet [202]	0.3807	0.5689	0.5539	0.3494
<b>Ours</b>	HKMNet	0.4298	0.6518	0.6043	0.4008
	HKMNet + $\mathcal{H}_{\mathcal{E}}$	<b>0.4706</b>	<b>0.6684</b>	<b>0.6433</b>	<b>0.4379</b>



(a) ROC curves (Secondary-Kin vs Non-kin) trained on 7 kin types (b) ROC curves (Primary-Kin vs Unlabeled) trained on 7 kin types

Figure 40: ROC curves of the different methods for FIW.

For a fair comparison, networks are used with the same backbone [225]. AttentionNet + *Softmax* is created following [137]. It uses kin types in the training set directly as classes. Moreover, two variations are added: AttentionNet+*Softmax*\* and AttentionNet\*. AttentionNet+*Softmax*\* utilizes non-kin samples during training. AttentionNet\* uses the same auto-annotated label as our HKMNet model.

We also compare our method with related kinship recognition methods. CNN-point [240] is selected as a representative of deep learning methods for kinship verification. JLNet is a pairwise-based kinship identification model. It utilizes kinship verification ensembles to enhance kinship identification performance. CAC, ARPL, and JLNet use kin-types as their training classes following their training set in the experiments. CNN-point uses the same auto-annotated label as our HKMNet.

### Results and Discussion

Accuracy measures the matching of correct categories. It corresponds to the ability in positive samples among different models. AUROC (Primary-Kin vs unlabeled) corresponds to the ability to differentiate known labeled and unlabeled pairs where unknown samples contain both unknown-kin-related and unknown-non-kin pairs. AUROC (Secondary-Kin vs Non-Kin) measures the ability to retrieve unlabeled but kin-related pairs from all



Table 25: Precision, Recall and F1 of 5-cross-validation for each testing category of different methods on the FIW dataset. Here, 7 kin types (F-D, F-S, M-D, M-S, B-B, B-S, S-S) within the Primary-Kin category are used for training.

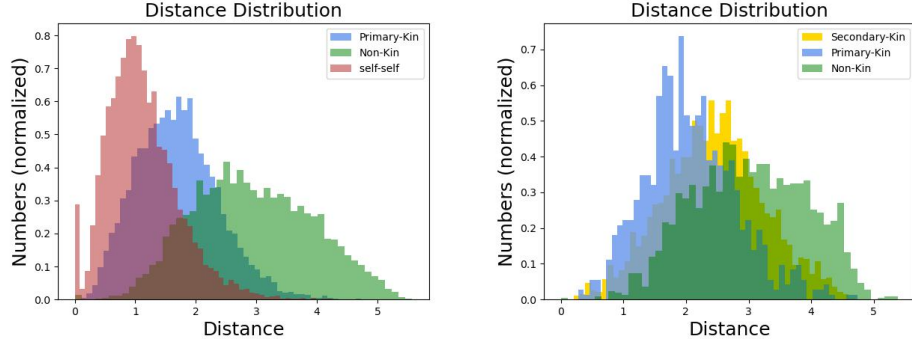
Methods		Primary-Kin			Secondary-Kin			Non-kin			Total F1-Macro
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	
Same-backbone	AttentionNet+ <i>Softmax</i> * [225]	0.4490	0.4826	0.4538	0.2609	0.2824	0.2243	0.3711	0.4565	0.3682	0.3488
	AttentionNet+ <i>Softmax</i> [225]	0.3373	0.3992	0.3344	0.2401	0.1546	0.1783	0.2966	0.4390	0.3379	0.2835
Open set	CAC [137]	0.3486	0.2696	0.2537	0.0667	0.0171	0.0272	0.2941	0.7118	0.3982	0.2264
	ARPL [26, 27]	0.5757	0.3362	0.4245	0.3206	0.7335	0.4462	0.2642	0.0745	0.1162	0.2590
kinship Related methods	CNN-point [240]	0.4777	0.5162	0.4878	0.3399	0.1905	0.2215	0.3850	0.5654	0.4403	0.3832
	AttentionNet* [225]	0.3436	0.2778	0.2397	0.1411	0.2264	0.1393	0.2612	0.5037	0.2935	0.2242
	JLNet [202]	0.4142	0.4695	0.4163	0.3775	0.2961	0.2743	0.3594	0.3859	0.3576	0.3134
<b>Ours</b>	HKMNet	0.4617	0.5883	0.4904	0.3790	0.2465	0.2703	0.4193	0.4742	0.4418	0.4008
	HKMNet + $\mathcal{H}_E$	0.4802	0.5717	0.5122	0.3934	0.3627	0.3689	0.4918	0.4119	0.4327	<b>0.4379</b>

unlabeled samples. F1 score computes the average F1 scores for three categories. Furthermore, precision, recall, and F1 scores are calculated for each category. We conduct 5-cross validation for the final results. Our HKMNet outperforms other methods for all metrics (e.g. distribution of distances and ROC curves).

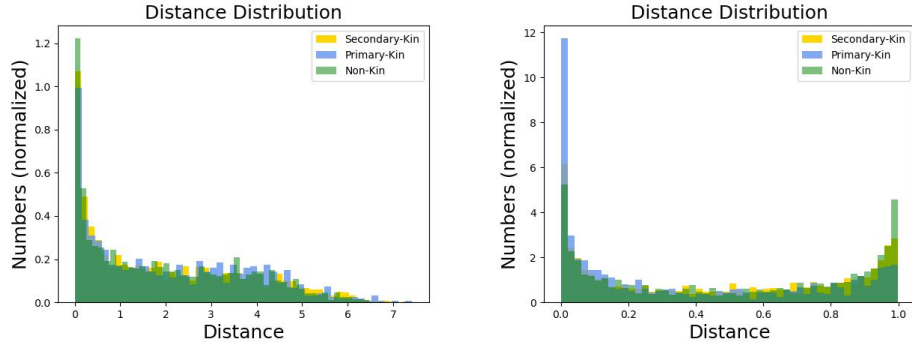
**COMPARISON WITH SOTA OPEN-SET METHODS** Table 24 shows a comparison of our method with SOTAs in OSR. Our approach, HKMNet, outperforms ARPL [26] and CAC [137] in all the four metrics. Note that  $\mathcal{T}_1$  and  $\mathcal{T}_2$  are derived from two ROC curves based on the Youden index [234]. If  $\mathcal{T}_1 > \mathcal{T}_2$ , there will be no optimal threshold for  $S$ . Then, Precision, Recall, and F1 will be set to 0. According to the results in Table 25, CAC and ARLP perform poorly in separating Non-kin type and unlabeled but kin-related pairs. ARLP obtains the lowest recall and F1 for Non-Kin CAC obtains the lowest recall and F1 for Secondary-Kin, which means it does not have the capability to distinguish and mine the unlabeled but kin-related pairs. Instead, our proposed model obtains more balanced scores.

**COMPARISON WITH SOTA KINSHIP RECOGNITION METHODS** CNN-Point, AttentionNet\*, and JLNet are kinship recognition related methods and are specifically designed to extract kinship related features. The results show that CNN-Point and JLNet perform, in general, better than CAC and ARPL. AttentionNet\* utilizes the same feature backbone and label as HKMNet(ours). The results between HKMNet and AttentionNet\* indicate the feasibility of our proposed methods. Figure 41 illustrates the distributions of distances for HKMNet (ours), CAC and JLNet on the testing dataset. For CAC, the distance distributions of three testing categories are overlapping. As for JLNet, it can distinguish the known Primary-Kin and Non-kin categories. However, the Secondary-Kin category does not follow a uniform distribution. In contrast, our method is able to properly separate the three categories. Although the Secondary-Kin is unknown to our HKMNet, our model can still form a unified similarity distribution for the Secondary-Kin category.

**COMPARISON WITH THE SAME BACKBONE CNNs** The results are shown in Table 24. Our model outperforms all (same) backbone models. As for precision, recall, and F1 among Secondary-Kin and Non-kin in Table 25, AttentionNet+*Softmax* performs in an unbalanced way. When comparing metric AUROC between AttentionNet+*Softmax*



(a) Distribution of distances of HKMNet on the training dataset (b) Distribution of distances of HKMNet on the testing dataset.



(c) Distribution of distances of CAC on the testing dataset. (d) Distribution of distances of JLNet on the testing dataset.

Figure 41: Distribution of distances of HKMNet (ours), CAC and JLNet.

Table 26: Ablation Study of HKMNet on FIW dataset.

$\tanh$	$fc$ share weights	$\mathcal{H}_E$	Accuracy	AUROC	AUROC	F1
				Primary-Kin vs Unlabeled	Secondary-Kin vs Non-Kin	
	✓		0.4193	0.6486	0.6066	0.3772
✓			0.4444	0.6472	0.6098	0.3885
✓	✓		0.4298	0.6518	0.6043	0.4008
✓	✓	✓	<b>0.4706</b>	<b>0.6684</b>	<b>0.6433</b>	<b>0.4379</b>

and AttentionNet+*Softmax*\*, it is shown that adding non-kin label information improves the capability of the AttentionNet+*Softmax*\* to distinguish different categories. Even though using the same auto-annotated labels, our HKMNet still outperforms AttentionNet\*. This means that our proposed Hierarchical Kinship Triplet Loss is more beneficial than the Softmax+Cross-Entropy loss.

### Ablation Study

An ablation study is conducted. When removing the  $\tanh$  activation function, the performance drops.  $\tanh$  activation function enforces the feature values to a limited scale, which helps the model to obtain a better representation. When the  $fc$  layers do not

Known											
Test Pairs											
Ground Truth	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin
AttentionNet+softmax*	Secondary-Kin	Secondary-Kin	Non-kin	Non-kin	Secondary-Kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin
AttentionNet+softmax	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Secondary-Kin	Secondary-Kin	Non-kin	Secondary-Kin	Non-kin	Non-kin
AttentionNet*	Secondary-Kin	Non-kin	Secondary-Kin	Secondary-Kin	Secondary-Kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin
CAC	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin
ARPL	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin
CNN-point	Secondary-Kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Secondary-Kin	Non-kin	Non-kin
JLNet	Non-kin	Non-kin	Non-kin	Non-kin	Secondary-Kin	Secondary-Kin	Secondary-Kin	Non-kin	Secondary-Kin	Secondary-Kin	Non-kin
HKMNet (ours)	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin

Unknown (Unlabeled)											
Test Pairs											
Ground Truth	Secondary-Kin	Secondary-Kin	Secondary-Kin	Secondary-Kin	Secondary-Kin	Secondary-Kin	Non-Kin	Non-Kin	Non-Kin	Non-Kin	Non-Kin
AttentionNet+softmax*	Primary-Kin	Primary-Kin	Non-kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Secondary-Kin
AttentionNet+softmax	Primary-Kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Secondary-Kin
AttentionNet*	Non-kin	Primary-Kin	Primary-Kin	Non-kin	Primary-Kin	Non-kin	Secondary-Kin	Secondary-Kin	Primary-Kin	Secondary-Kin	Secondary-Kin
CAC	Primary-Kin	Non-kin	Primary-Kin	Non-kin	Non-kin	Non-kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin
ARPL	Non-kin	Non-kin	Non-kin	Primary-Kin	Non-kin	Primary-Kin	Secondary-Kin	Primary-Kin	Secondary-Kin	Secondary-Kin	Secondary-Kin
CNN-point	Primary-Kin	Non-kin	Non-kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Secondary-Kin	Secondary-Kin	Secondary-Kin	Primary-Kin
JLNet	Primary-Kin	Non-kin	Non-kin	Primary-Kin	Primary-Kin	Non-kin	Primary-Kin	Primary-Kin	Primary-Kin	Primary-Kin	Secondary-Kin
HKMNet (ours)	Secondary-Kin	Secondary-Kin	Secondary-Kin	Secondary-Kin	Secondary-Kin	Secondary-Kin	Non-kin	Non-kin	Non-kin	Non-kin	Non-kin

Figure 42: Qualitative results of different methods. The testing set consists of three kinship categories: "Primary-Kin", "Secondary-Kin" and "Non-Kin". Here, "Secondary-Kin" and "Non-Kin" are unknown and unlabeled pairs in the open set.

share weights, it is difficult for the method to learn a uniform projection for the different extracted features. When mining hard samples during the training process, the AUROC score increases. The Accuracy and average F1 are also increased. The hard sample mining process can be a good process to improve the performance.

### Qualitative Results

Qualitative results are shown in Figure 42. "Secondary-Kin" pairs are unknown and unlabeled in the open set. They are easy to be taken as "Primary-Kin" or "Non-Kin". Compared to other methods, our model (HKMNet) is able to correctly recognize the testing pairs in the "Primary-Kin", "Secondary-Kin", and "Non-Kin" categories.

## 5.6 CONCLUSION

A method has been proposed to determine family relationships and their corresponding degrees of kinship in an open set environment. It is pairwise-based and is able to exploit mutual information from positive pairs in a hierarchical way. Experiments and an ablation study show that our method outperforms the compared methods. Our model is able to properly separate kinship categories, and generates uniform similarity distributions.

Our approach has the following benefits for related kinship research: (1) pairs may exist (i) without any genetic relationship or (ii) with unlabeled kin-relationships in open-set collections. This may lead to (hard) negative samples possibly affecting the performance of close-set trained models in a negative way. Our method determines general kinship similarities of potential pairs to identify (hard) non-kin samples. This may yield cleaner close-set collections for kinship-related tasks, (2) current publicly available datasets may contain kin-related pairs without labels (*i.e.* kinship degree). Our approach is able to measure similarities for these unlabeled but kin-related samples yielding enhanced and larger kinship datasets.

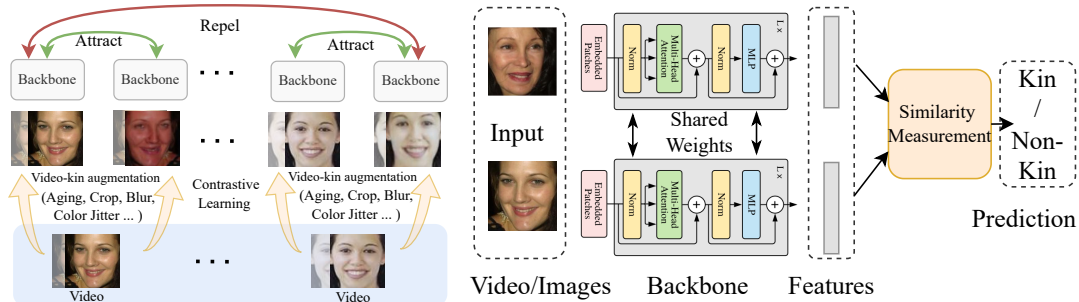
## KINSHIP VERIFICATION IN VIDEOS USING SEMI-SUPERVISED LEARNING

### 6.1 INTRODUCTION

Vision-based kinship verification [130, 165, 166] aims to determine whether faces in images are kin or non-kin. It is an important task in computer vision as there are many applications such as media analysis [20, 43, 95, 236], missing children searching [217] and social behavior analysis [64, 93, 146, 240]. Current kinship datasets, especially video datasets, are relatively small because manually collecting facial images/videos with labels is difficult and tedious. Hence, learning features for kinship verification with a limited number of samples is a problem.

To deal with small datasets, in this chapter, transfer learning is used for kinship verification. Large-scale face datasets (not intended for the kinship verification task) are used to extract facial features. Learning kinship verification using pre-training without kinship annotations has not been studied before. In contrast to a typical classification task based on a single subject, kinship information is represented by the relationship between two subjects. Then, the simulation of the kinship distribution through unlabeled external face data becomes important.

Therefore, in this chapter, we propose a kinship-oriented augmentation method (Video-kin augmentation) for kinship verification in videos. As shown in Figure 43, during the pre-training stage, the original videos are augmented using different styles. A series of frames are augmented through age transformation and face deformation. These frames



(a) Pretraining on external large scale face datasets. (b) Training/testing on kinship dataset based on pretrained backbone.

Figure 43: Pipeline of proposed Kinship-transformer.

are then used to form visually similar video pairs, where the appearance is similar, but the age differs slightly. The augmentation enables the model to form feature representations similar to the kinship distributions. The proposed method is divided into three steps. Firstly, the proposed Kinship-transformer is pre-trained on large scale face datasets (*e.g.* YouTube Face Database [211]) which are designed for tasks other than kinship verification, *i.e.* no kinship labels. Then, different strategies for data augmentation are used. Facial and dynamic feature representations are learned through contrastive learning. Finally, the pre-trained Kinship-transformer is fine-tuned on (small) video kinship datasets.

The contributions of this chapter are summarized as follows:

- A kinship-oriented augmentation method, Video-kin augmentation, is proposed to enable the model to learn kinship-like distributions based on large face video datasets.
- Video transformers are proposed for the kinship video verification task.
- The proposed framework can be pre-trained on large face datasets without kinship annotations.
- Experiments show that the proposed method outperforms existing convolutional neural networks.

## 6.2 RELATED WORK

### 6.2.1 Kinship Verification

Kinship verification in computer vision is considered as a binary classification problem. It determines whether two or more persons, represented by image/video pairs of their faces, are kin related or not [165]. Different methods are proposed in the field of kinship verification [31, 50, 78, 112–114, 130, 219, 228, 231, 251]. Among these methods, image-based kinship verification methods are mostly studied. In contrast, video-based kinship verification has received less attention. Dibeklioglu *et al.* [50] use a video dataset for kinship verification. Both statistical and dynamic features are extracted. Boutellaa *et al.* [19] use shallow spatio-temporal learning for kinship verification. Yan *et al.* [226] collect a new dataset (KFVW) and evaluate different metric learning methods. Dong *et al.* [53] aggregate multiple visual features using multi-modal knowledge and design an adaptive feature fusion mechanism. The method is used in a self-supervised way, assuming that each sample pair is a distinct class of its own. A memory bank (Moco [82]) and noise-contrastive estimation (InfoNCE [189]) are utilized. Zhang *et al.* [239] propose a linear combination model to measure the similarity between parents and children in an unsupervised manner. However, this method is based on tri-subject verification and is not suitable for the common task of bi-subject kinship verification. In contrast to the above mentioned methods, in this chapter, the aim is to learn features for video-based kinship verification using pre-training without kinship annotations in a semi-supervised way.

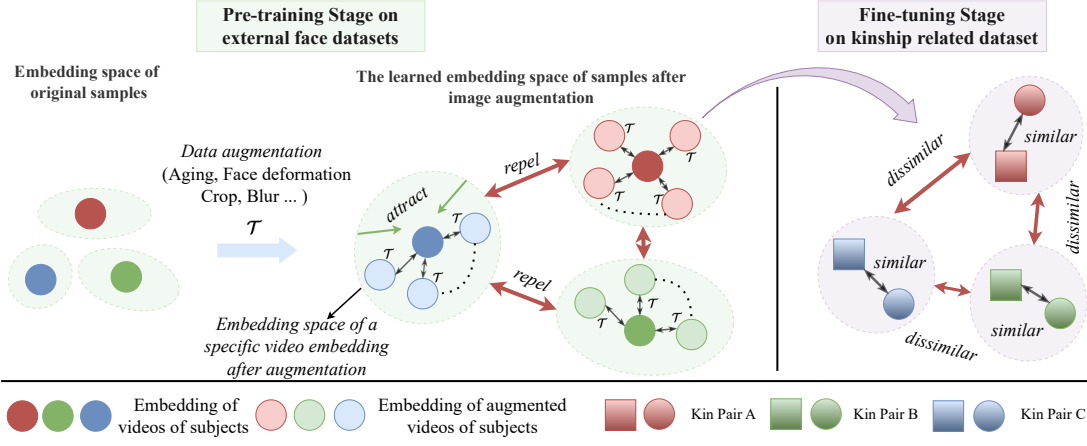


Figure 44: Feature space of embeddings of videos after Video-kin augmentation  $\mathcal{T}$ .

### 6.2.2 Transformer

Dosovitskiy *et al.* introduce the Vision Transformer (ViT) for computer vision tasks [55]. Several attempts are proposed to improve the performance such as Pyramid Vision Transformer (PVT) [201], Swin Transformer [122], Twins [34], DeiT [187], and DETR [22]. DETR focuses on end-to-end object detection with transformers. PVT utilizes a progressive shrinking pyramid to reduce the computation. Swin Transformer merges image patches to build hierarchical feature maps. However, DETR, PVT and Swin Transformer are data agnostic. Instead, Deformable Attention Transformer (DAT) [222] trains the transformer in a data-dependent way by using a deformable self-attention module.

The success of transformers also inspires researchers to study video-based recognition tasks. The Video Transformer Network (VTN) [144] uses a temporal attention mechanism for video recognition. MViT [60] uses multiscale ViT and learns spatio-temporal information. Video Swin Transformer (SWT) studies the spatio-temporal locality and extends the Swin transformer [124] from 2d-shifted windows to 3d-shifted windows.

### 6.2.3 Pre-training without Annotation

Manually collecting face images/videos with kinship labels is difficult and tedious. Hence, current annotated kinship datasets are relatively small. Semi-supervised learning (*i.e.* face images without kinship relations) can alleviate this problem through pre-trained feature extraction [132] from unlabeled data [90]. Combining unsupervised feature learning [90] and transfer learning [132] using large-scale face datasets can be beneficial. Contrastive learning aims to discriminate between positive (similar) and negative (diverse) samples by similarity measurements such as SimCLR [28], SwAV [23], and MoCo [82]. For generative learning, Generative Adversarial Networks (GANs) [74], and auto-encoders [191] are often used. Recently, He *et al.* propose masked autoencoders (MAE) [81] using a transformer architecture.

### 6.3 METHOD

#### 6.3.1 Problem Formulation

As discussed in the Section. 6.2, kinship verification is a binary classification task and determines whether target images are kin or not kin. Kinship verification [114] can be formalized as follows. Consider  $\mathcal{P} = \{(\mathbf{x}_i^a, \mathbf{x}_i^b) \mid i = 1, 2, \dots, N\}$  as the training set of image pairs containing kin relationships for each kin-type, where  $N$  is the number of positive pairs.  $\mathbf{x}_i^a$  and  $\mathbf{x}_i^b$  are parent image/video and children image/video, respectively. Then, the negative training set is denoted by  $\mathcal{N} = \{(\mathbf{x}_i^a, \mathbf{x}_i^b) \mid i = 1, 2, \dots, N, i \neq j\}$ , representing image pairs without kinship. To verify kin types, a binary classifier  $f(\cdot)$  is used and is formulated by:

$$\mathbf{y} = f(g(\mathbf{x}_i^a, \mathbf{x}_j^b)), \mathbf{y} \in \{0, 1\}, \quad (31)$$

where 1 represents kin and 0 represents non-kin. In this chapter,  $g(\cdot)$  represents the transformer architecture and  $f(\cdot)$  represents the cosine similarity.

Because existing kinship video dataset are limited, semi-supervised learning is used through pre-trained feature extraction from face data without kinship relations. Therefore, our method consists of a *pre-training* and a *fine-tuning stage*, see Figure 43. In the first stage, the network (Kinship-transformer) is pre-trained on (external) face datasets without kinship relations. In the second stage, the pre-trained model is fine-tuned on kinship-related datasets.

#### 6.3.2 Feature Learning through Video-kin Augmentation

Different from the standard single-subject task, the kinship relation corresponds to the relation between two subjects. For instance, parent-child pairs share visual similarities and differences in their ages. These features are utilized during pre-training to improve the model's performance. The appearance of the same person can vary due to both intrinsic (*e.g.* aging, expression) and extrinsic (*e.g.* saturation, contrast) changes. In this chapter, a wider range of features is formed through video augmentation. Figure 44 shows the feature learning process of our pre-training stage using augmentation.

In the pre-training stage, in addition to classical augmentation (such as changes in contrast, flipping and color), the original video is further augmented using appearance changes caused by variations in aging. As a result, the feature embeddings of one specific video are formed into multiple adjacent features. Such hidden space of the specific video embeddings should be clustered. Therefore, following the strategy of SimCLR [28], augmented video pairs are generated using the video-kin augmentation methods  $\mathcal{T}$  at each epoch, and use the InfoNCE loss [148] to repel negative and attract positive samples. After pre-training without kinship relations, the network learns the hidden embedding space of the external face datasets and forms the capacity to discriminate similar and diverse samples. In contrast to SimCLR, we extend the data augmentation process to videos and do not use a projection head during the unsupervised kinship pre-training



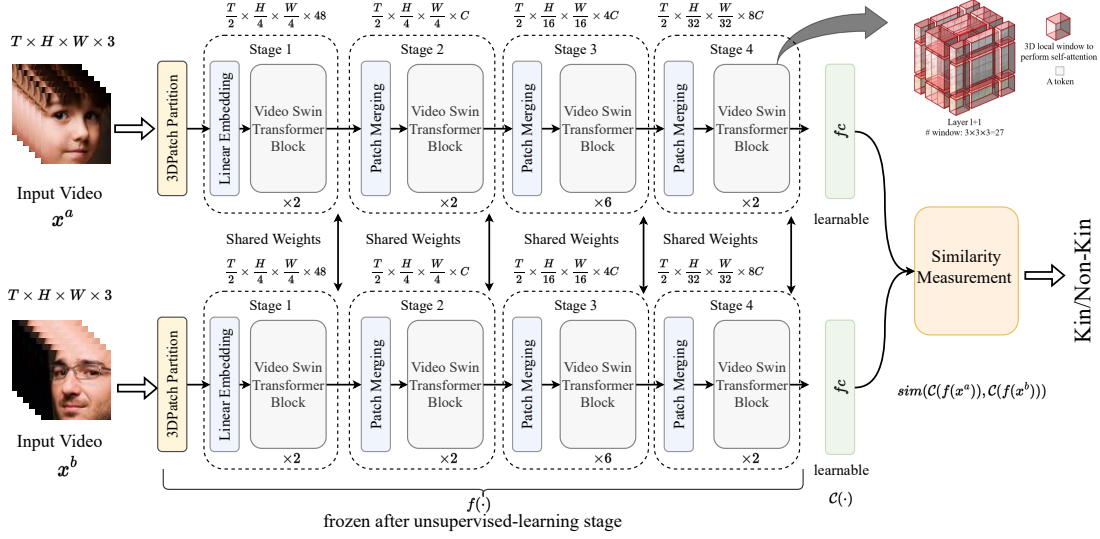


Figure 45: Architecture of our proposed Kinship-transformer-VST (KT-VST) with the VST [124] backbone.

stage. Experiments show that without the projection head, our network achieves better results.

During the fine-tuning stage, the pre-trained network is trained on kinship datasets. Positive kinship pairs are inherently similar to each other. The network is fine-tuned to form meaningful kinship feature distributions based on prior feature distribution representations.

### 6.3.3 Kinship-transformer

Vision transformers can be used as the backbone of our proposed Kinship-transformer. In this chapter, the Video Swin Transformer (VST) [124] is used. As shown in Figure 45, the pre-trained transformer is composed of a two-branch architecture followed by a similarity measurement module.

Given a 3D window with size  $P \times M \times M$ , the input videos  $\mathbf{x}^a \in \mathbb{R}^{T \times H \times W \times C}$ ,  $\mathbf{x}^b \in \mathbb{R}^{T \times H \times W \times C}$  are partitioned into  $\frac{T}{P} \times \frac{H}{M} \times \frac{W}{M}$  non-overlapping 3D windows. The partitioned windows are shifted along the time, height and width axes by  $\frac{P}{2} \times \frac{M}{2} \times \frac{M}{2}$ .

Then, a block of the transformer is formulated by:

$$\begin{aligned}
 \mathbf{z}'_\ell &= \text{3DW-MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, & \ell &= 1 \dots L, \\
 \mathbf{z}_\ell &= \text{FFN}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, & \ell &= 1 \dots L, \\
 \mathbf{z}'_{\ell+1} &= \text{3DSW-MSA}(\text{LN}(\mathbf{z}_\ell)) + \mathbf{z}_\ell, & \ell &= 1 \dots L, \\
 \mathbf{z}_{\ell+1} &= \text{FFN}(\text{LN}(\mathbf{z}'_{\ell+1})) + \mathbf{z}'_{\ell+1}, & \ell &= 1 \dots L,
 \end{aligned} \tag{32}$$

where  $\text{LN}$  is the Layernorm and 3DW-MSA is the 3D window multi-headed self-attention. 3DSW-MSA is the 3D shifted window multi-headed self-attention. The output embeddings of image  $\mathbf{x}^a$  and  $\mathbf{x}^b$  are denoted by  $\mathbf{z}_o^a$  and  $\mathbf{z}_o^b$ . Finally,  $\mathbf{z}_o^a$  and  $\mathbf{z}_o^b$  are projected by

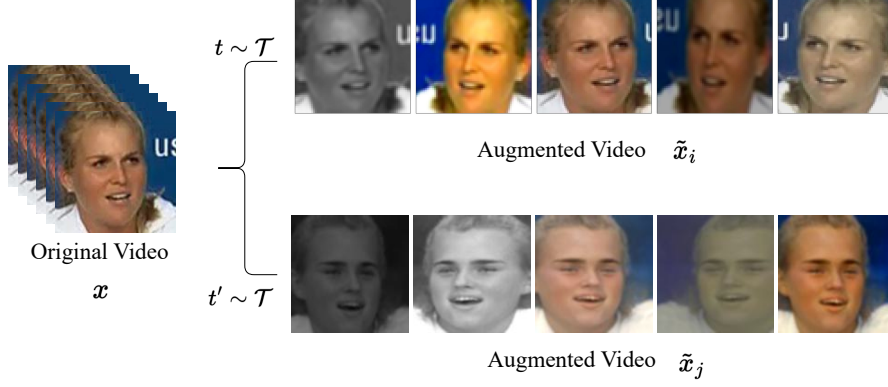


Figure 46: Illustration of video data augmentation.

the  $fc$  layer and used by the similarity measurement module ( $sim$ ) for the final kinship verification:

$$y = sim(C(z_o^a), C(z_o^b)), \quad (33)$$

where  $C$  is the  $fc$  layer projection process and  $sim$  is the similarity measurement (cosine similarity). During training on the kinship dataset, the pre-trained transformers are frozen, and only the  $fc$  layer is learnable.

#### 6.3.4 Pre-training of Kinship-transformer

The semi-supervised pre-training of our network aims to learn face feature representations between videos and videos' augmentations. To this end, pair-wise video data augmentation with contrastive learning is utilized on face datasets (*e.g.* YouTube Face Database) without kinship relations.

##### Video augmentation

As shown in Figure 46, a training video  $x_i$  is augmented by two random data augmentations from the set of augmentations ( $t \sim \mathcal{T}$  and  $t' \sim \mathcal{T}$ ), which leads to two newly augmented videos  $\tilde{x}_i$  and  $\tilde{x}_j$ . Since these two augmented videos are computed for the same videos, they are taken as a positive pair. The Kinship-transformer (KT) is denoted by  $f(\cdot)$ . The feature representation of the augmented video  $\tilde{x}_i$  after kinship transformer is given by  $\tilde{z}_i = f(\tilde{x}_i)$ .

##### InfoNCE Loss

The InfoNCE loss [28] is used as the training loss. Assuming  $N$  videos in a mini-batch,  $2N$  augmented videos are generated after data augmentation. Consequently, for one positive pair, the other  $2(N - 1)$  augmented videos are negative. The loss function is given by:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j) / \tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k) / \tau)}, \quad (34)$$

where  $\text{sim}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\top \mathbf{z}_j / \|\mathbf{z}_i\| \|\mathbf{z}_j\|$  is a cosine similarity function and  $\tau$  is defined as a parameter.  $\mathbf{1}_{[k \neq i]} \in \{0, 1\}$  indicates 1 if  $k \neq i$ .

### 6.3.5 Fine-tuning of Kinship-transformer on Kinship Related Datasets

During fine-tuning of the Kinship-transformer,  $f(\cdot)$  is frozen. The kinship datasets are used as training sets. A learnable  $fc$  layer is designed to map the facial representation to a kinship salient manifold. Then, the cosine similarity function is used as the final result. During training, the MSE loss is utilized as the training loss.

### 6.3.6 Similarity during Testing

During the evaluation process, the cosine similarity is utilized as the final output of the Kinship-transformer. A threshold is used to predict whether the inputs are positive or not. The prediction is defined by:

$$\text{decision} = \begin{cases} 1 & \text{if } \text{sim}(C(f(x^a)), C(f(x^b))) > \theta \\ 0 & \text{if } \text{sim}(C(f(x^a)), C(f(x^b))) \leq \theta \end{cases}, \quad (35)$$

where  $\theta$  is the threshold for similarity.

## 6.4 EXPERIMENTS

Table 27: Different methods on the Nemo-Kinship dataset. Here, with labels means pre-training with face labels (identification labels), without labels means pre-training without any other labels.

Model	Type	FD $\uparrow$	FS $\uparrow$	MD $\uparrow$	MS $\uparrow$	BB $\uparrow$	BS $\uparrow$	SS $\uparrow$	Mean $\uparrow$
Sphereface-baseline [165]	fully supervised	0.524	0.541	0.561	0.555	0.594	0.581	0.547	0.557
Vuvko [175]	fully supervised	0.775	0.849	0.777	0.734	0.817	0.761	0.761	<b>0.772</b>
DEEP+Shallow [19]	fully supervised	0.583	0.567	0.576	0.571	0.467	0.700	0.533	0.571
KT-IResnet (ours)	Semi-supervised	0.527	0.517	0.580	0.567	0.564	0.505	0.540	0.543
KT-ViT (ours)	Semi-supervised	0.418	0.518	0.581	0.553	0.540	0.581	0.529	0.531
KT-VST (ours)	Semi-supervised	0.583	0.563	0.631	0.594	0.667	0.633	0.600	<b>0.610</b>

### 6.4.1 Datasets

Two face datasets are used for pre-training without using any of their labels: MS1M-retinaface (Lightweight Face Recognition Challenge & Workshop (ICCV 2019)) [79] and YouTube Face Database (YTB) [211]. MS1M-retinaface is used during pre-training for the image-based backbones of our Kinship-transformer. MS1M-retinaface dataset is cleaned from MS1M [79]. As mentioned on the official website<sup>1</sup>, all face images are pre-processed to size 112x112 by five facial landmarks predicted by RetinaFace [46]. In total, there are 5.1M images with 93K identities. MS1M-retinaface is used during pretraining by the image-based backbones. YouTube Face Database (YTB) is a video-based facial dataset for facial recognition and related tasks. YTB consists of 3425 videos

<sup>1</sup> <https://ibug.doc.ic.ac.uk/resources/lightweight-face-recognition-challenge-workshop/>

with 1595 identities. All these videos are collected from YouTube without kinship relations. It is used during semi-supervised pre-training of the video-based backbones (Kinship-transformer).

Several datasets are selected for our video-based kinship verification task. Nemo-kinship dataset is used for video-based kinship verification <sup>2</sup>. We select 7 types: Father-Daughter (FD), Father-Son (FS), Mother-Daughter (MD), Mother-Son (MS), Sister-Brother (SB), Sister-Sister (SS), and Brother-Brother (BB). KinFaceW-I and KinFaceW-II [130] are also used. KinFaceW-I&II are image-based kinship verification datasets. In KinFaceW-I, kinship pairs are collected from different pictures. In KinFaceW-II, all pairs are obtained from the same picture. These pictures are unconstrained in terms of pose, lighting, background, expression, age, ethnicity, and partial occlusion. There are four types of kinship relations for these two datasets. In KinFaceW-I, there are 156 pairs of F-S, 134 pairs of F-D, 116 pairs of M-D, and 127 pairs of M-S. Meanwhile, in KinFaceW-II, there are 250 pairs of pictures for each kinship relation.

#### 6.4.2 Implementation Details

In the experiments, three networks are used as backbones of the Kinship-transformer: Video Swin Transformer (VST) [124], Vision Transformer [55] and IResNet [59].

For the Kinship-transformer with the VST backbone (KT-VST), the embedding dimension is set to 48. The depths are [2, 2, 6, 2]. We use the patch size of 2 in the first block and with a size of 4 in the last two blocks. The window size is set to [8, 7, 7]. For the Kinship-transformer with ViT backbone (KT-ViT), the number of layers is 20. The number of heads is 8. The hidden size is 512. The patch size is 8 and  $N = 64$ . The position embeddings (not relative position) are the learnable parameters initialized following a normal distribution. For the Kinship Kinship-transformer with the IResNet backbone (KT-IResNet), the layers are set to 100. The output feature is 128. During the training, the learning rate is  $1e^{-3}$ .

**PRETRAINING ON FACE DATASETS** During the pre-training stage, KT-VST is trained on YTB. Before training, all videos in YTB are extracted, and labels (e.g. identity information) are ignored. During training, each video in the mini-batch is sent to the augmentation operators, which are randomly selected from the augmentation lists. Each video forms two augmented videos. We denote the videos augmented from the same video as positive samples. All others are used as negative samples. The InfoNCE loss is utilized. The augmentation lists are Horizontal Flip, Random Resized Crop, Color Jitter, Gaussian Blur, Random Gray Scale, aging generation [208], and face deformation. The epoch is set to 100, and the batch size is 40. The learning rate is  $1e^{-4}$ . For the image-based backbones (KT-ViT and KT-IResNet), MS1M-retinaface is used as the pre-training set. All images are extracted and stored together.

<sup>2</sup> <https://www.nemosciencemuseum.nl/nl/wat-is-er-te-doen/activiteiten/science-live/>

**FINE-TUNING ON KINSHIP DATASETS** During the fine-tuning stage, Nemo-kinship is utilized as training and testing data for video-based kinship verification. The same augmentation method as in the pretraining stage is used. Horizontal flips and aging generation are utilized. As shown in Figure 45, the backbone of the Kinship-transformer is frozen after pre-training. The backbone extracts features of two input images/videos. A learnable  $fc$  layer  $C(\cdot)$  is utilized to map features to a kinship-related manifold. For KT-VST, the feature size of each video is 374 and the feature size after mapping is 64. Two mapped features are normalized and compared by the cosine similarity function. In this stage, the label information is used. During the training on Nemo-Kinship dataset, the similarity output is supervised by 0 or 1 label based on  $MSE_{loss}$ . The batch size is 64. The epoch for each cross-validation is 60. The learning rate is set to  $1e^{-4}$ . *AdamW* is used as the optimizer. For KinFaceW-I and KinFaceW-II datasets, the image is taken as a short video with one frame. During training, the batch size is set to 64. For KT-Vit and KT-IResNet, the videos in Nemo-kinship are extracted into frames. The KT-Vit and KT-IResNet are trained in an image-based kinship verification manner.

### 6.4.3 Results

#### *Experiment results on Nemo-Kinship dataset*

Table 28: Results of different methods on the KinFaceW-I dataset.

Model	Type	FD $\uparrow$	FS $\uparrow$	MD $\uparrow$	MS $\uparrow$	Mean $\uparrow$
SMCNN [111]	fully supervised	0.750	0.750	0.722	0.687	0.727
CFT [56]	fully supervised	0.795	0.716	0.733	0.799	0.761
CFT* [56]	fully supervised	0.788	0.717	0.772	0.819	0.774
WGEML [117]	fully supervised	0.785	0.739	0.806	0.819	0.787
GKR [114]	fully supervised	0.795	0.732	0.780	0.862	0.792
DSMM [113]	fully supervised	0.767	0.817	0.890	0.823	<b>0.824</b>
KT-IResnet (ours)	Semi-supervised	0.538	0.587	0.605	0.564	0.573
KT-ViT (ours)	Semi-supervised	0.635	0.651	0.666	0.608	0.640
KT-VST (ours)	Semi-supervised	0.657	0.682	0.707	0.634	<b>0.670</b>

Since the proposed setting for kinship verification is novel, we can only compare the performance of our methods with different backbones. The results of different methods on YTB are shown in Table 27. The accuracy shows that our proposed KT-VST achieves 0.610 average accuracy on the video-based Nemo-kinship dataset. KT-IResNet and KT-ViT obtain 0.543 and 0.531 respectively. It shows that our video based KT-VST outperforms image-based methods (KT-ViT and KT-IResNet). Moreover, our model KT-VST shows better results on BB, BS and SS kin-type. The reason is that the age of the subjects in BB, BS and SS type are more similar. The similar age samples improve the feature representation after transfer learning.

Results are listed for current fully supervised kinship verification methods. For image-based methods, we extract frames ( $N$  frames for one video) of the videos and form  $N$  image-pairs. The average of  $N$  image-pairs is taken as the final result. As shown in

Table 29: Results of different methods on KinFaceW-II dataset.

Model	Type	FD $\uparrow$	FS $\uparrow$	MD $\uparrow$	MS $\uparrow$	Mean $\uparrow$
CFT [56]	fully supervised	0.754	0.688	0.774	0.778	0.759
SMCNN [111]	fully supervised	0.750	0.790	0.780	0.850	0.793
CFT* [56]	fully supervised	0.774	0.766	0.790	0.838	0.793
WGEML [117]	fully supervised	0.886	0.774	0.834	0.816	0.828
KML [228]	fully supervised	0.874	0.836	0.862	0.856	0.857
GKR [114]	fully supervised	0.908	0.860	0.912	0.944	0.906
DSMM [113]	fully supervised	0.898	0.926	0.958	0.936	<b>0.930</b>
KT-IResnet (ours)	Semi-supervised	0.650	0.712	0.680	0.718	0.690
KT-ViT (ours)	Semi-supervised	0.702	0.758	0.754	0.722	0.734
KT-VST (ours)	Semi-supervised	0.714	0.766	0.782	0.790	<b>0.763</b>

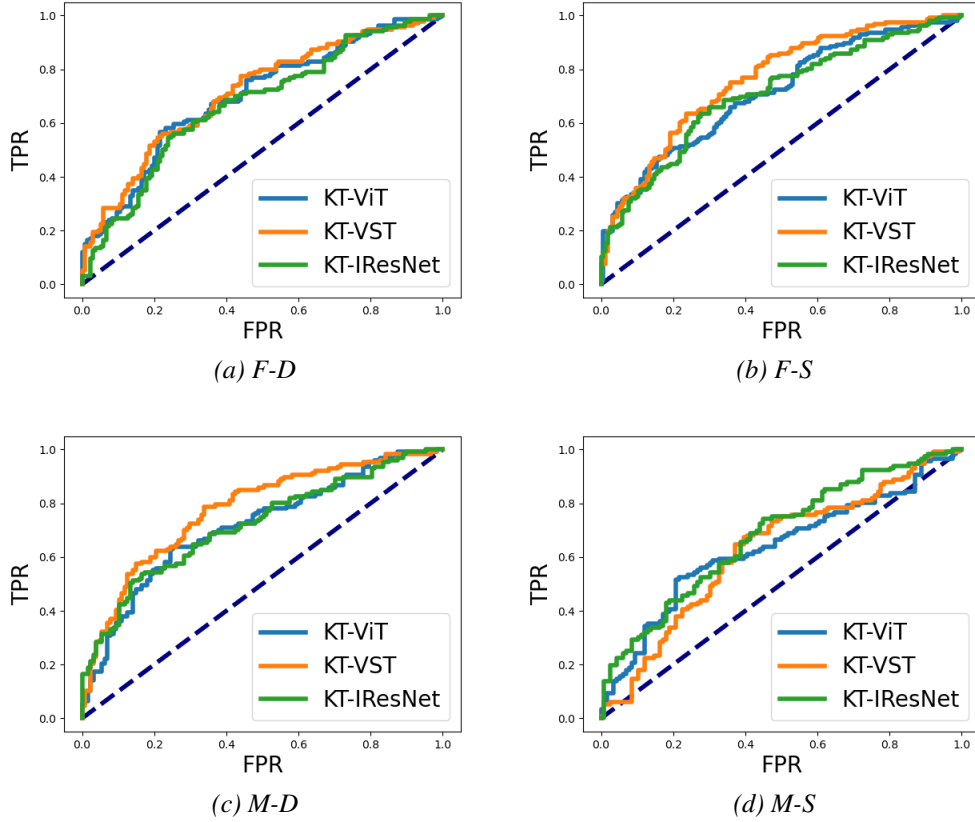


Figure 47: The ROC curves of Kinship-transformer for the fourth cross validation on KinFaceW-I dataset.

Table 27, Vuvko shows the best supervised performance. VuvKo is pretrained on MS-celeb-1M dataset using the ArcFace model. The model uses the off-the-shelf capability of the network and reaches the first place in RFIW2020 competition.

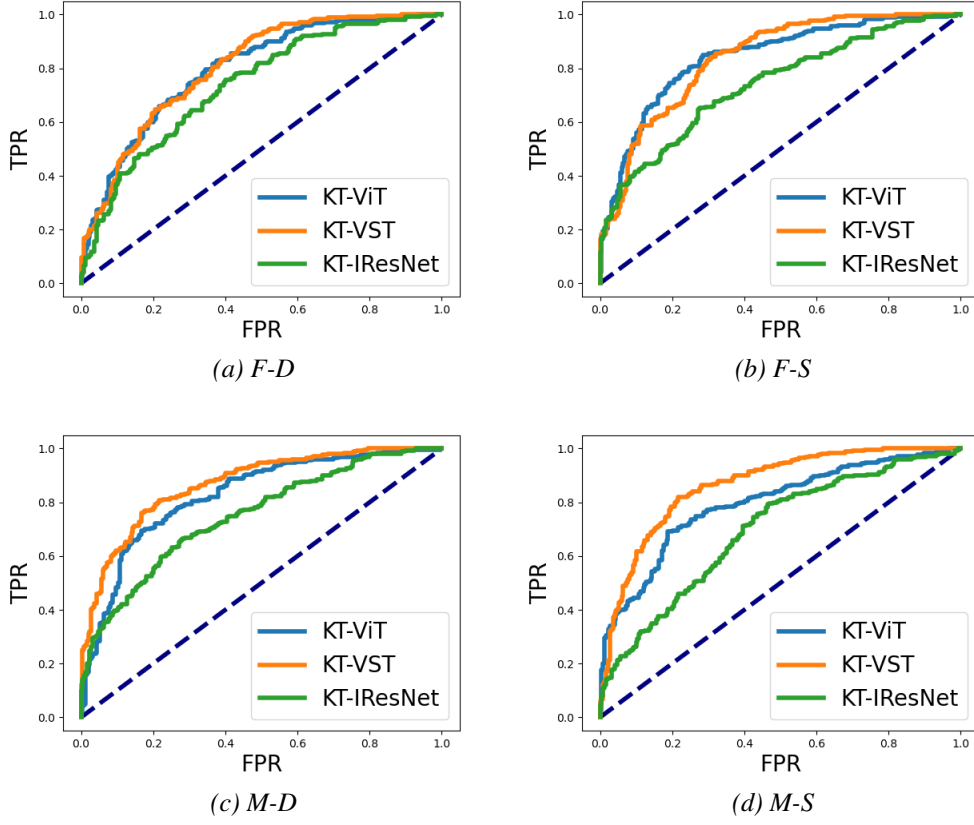


Figure 48: The ROC curves of Kinship-transformer for the fourth cross validation on KinFaceW-II dataset.

#### Results on KinFaceW-I and KinFaceW-II

Figure 47 and Figure 48 show the ROC curves of our proposed methods. In general, KT-VST achieves the best performance. Table 28 and Table 29 show the different methods on KinFaceW-I and KinFaceW-II datasets. Our KT-VST results in 0.670 on KinFaceW-I and 0.763 on KinFaceW-II. KT-VST outperforms the supervised method CFT on KinFaceW-II.

Feature distances are visualised for KT-VST on KinFaceW-II. Figure 49 show that our proposed KT-VST provides better separable feature distances on the MS type.

#### 6.4.4 Ablation Study

Table 30: Ablation studies on the Nemo-kinship dataset.

Data augmentation	Pretraining with annotations	Pretraining without annotations	Training on Nemo	FD↑	FS↑	MD↑	MS↑	BB↑	BS↑	SS↑	Mean↑
✓	-	-	✓	0.500	0.500	0.500	0.500	0.500	0.500	0.500	<b>0.500</b>
✓	✓	-	-	0.483	0.516	0.546	0.524	0.600	0.55	0.567	<b>0.541</b>
-	-	✓	✓	0.550	0.570	0.521	0.476	0.567	0.517	0.600	<b>0.543</b>
✓	-	✓	✓	0.583	0.563	0.631	0.594	0.667	0.633	0.600	<b>0.610</b>
✓	✓	-	✓	0.550	0.643	0.610	0.583	0.767	0.617	0.600	<b>0.624</b>

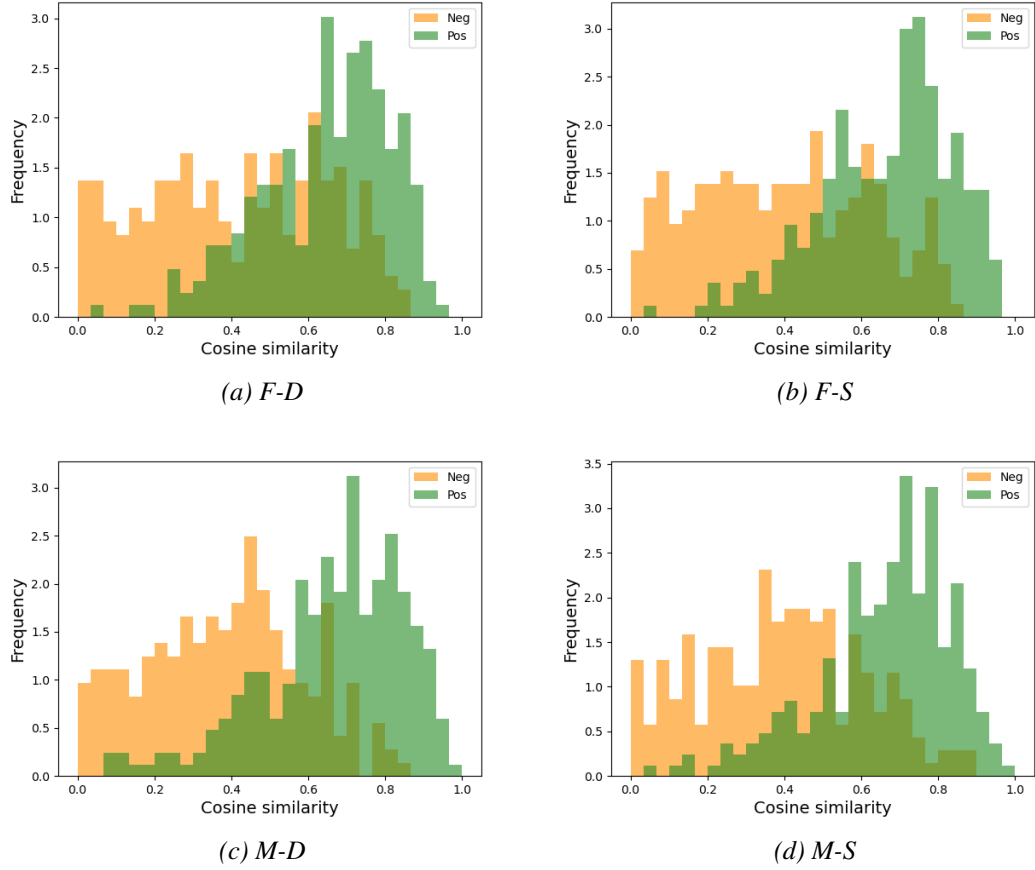


Figure 49: Feature distances of KT-VST for fourth cross validation on KinFaceW-II dataset. For visualization, we combine the cosine similarities of test pairs in all 5 cross validations.

Table 31: Ablation study using the projection head (SimCLR) and without (ours) during the semi-supervised pre-training stage.

Training Data	w/o projection head	FD $\uparrow$	FS $\uparrow$	MD $\uparrow$	MS $\uparrow$	BB $\uparrow$	BS $\uparrow$	SS $\uparrow$	Mean $\uparrow$
KinFaceW-I	-	0.623	0.599	0.639	0.556	-	-	-	0.604
KinFaceW-I	✓	0.657	0.682	0.707	0.634	-	-	-	<b>0.670</b>
KinFaceW-II	-	0.714	0.744	0.771	0.769	-	-	-	0.750
KinFaceW-II	✓	0.714	0.766	0.782	0.790	-	-	-	<b>0.763</b>
Nemo-kinship	-	0.617	0.533	0.532	0.644	0.600	0.533	0.500	0.566
Nemo-kinship	✓	0.583	0.563	0.631	0.594	0.667	0.633	0.600	<b>0.610</b>

Table 32: Comparisons with aging augmentation. DA corresponds to standard data augmentation.

Model	Pretraining	Fine-tuning	FD $\uparrow$	FS $\uparrow$	MD $\uparrow$	MS $\uparrow$	BB $\uparrow$	BS $\uparrow$	SS $\uparrow$	Mean $\uparrow$
KT-VST (ours)	DA	w/o DA	0.483	0.543	0.589	0.579	0.700	0.667	0.667	0.604
KT-VST (ours)	DA + aging	w/o DA	0.650	0.540	0.586	0.618	0.633	0.617	0.600	0.606
KT-VST (ours)	DA + aging	DA + aging	0.583	0.563	0.631	0.594	0.667	0.633	0.600	<b>0.610</b>

In this section, we discuss the ablation studies of our proposed network on the Nemo-kinship dataset. The related performances are listed in Table 30. There are five experiments. Firstly, the KT-VST is trained directly on the Nemo-Kinship dataset. The average accuracy is 0.5, which indicates that training of our network directly on the Nemo-kinship










<i>Test Pairs</i>	<i>KT-IResnet</i>	KT-ViT	KT-VST	Ground Truth
	Non-kin	Non-kin	Kin	Kin
	Non-kin	Non-kin	Kin	Kin
	Kin	Non-kin	Kin	Kin
	Non-kin	Non-kin	Kin	Kin
	Non-kin	Non-kin	Non-kin	Non-kin
	Kin	Non-kin	Kin	Non-kin
	Kin	Non-kin	Non-kin	Non-kin

Figure 50: Qualitative results for the first cross-validation of father-daughter type on KinFaceW-II dataset.

fails to converge. Secondly, the KT-VST is trained on YTB in a supervised manner and predicts the Nemo-kinship samples directly. The Nemo-kinship is not trained. Our KT-VST obtains 0.541. It shows that without fine-tuning on target kinship dataset, KT-VST fails to provide improved kinship representations. Thirdly, the KT-VST is trained without using Video-kin augmentation during the pretraining stage. The final accuracy drops by six percent. It shows the feasibility of our proposed Video-kin augmentation. Fourthly, the KT-VST is trained with the proposed pipeline. The KT-VST results in the best performance when pre-trained in a semi-supervised way. Finally, the KT-VST is trained in a supervised manner, and then fine-tuned on the Nemo-Kinship. The model achieves 0.624 among all the ablation studies. The table also shows that our proposed semi-supervised pipeline is competitive when compared to a fully supervised method. We conduct ablation studies to analyze the influence of the aging augmentation. Table 32 shows that aging augmentation improves the performance. We conduct extra ablation studies to analyze the performance of our method compared to SimCLR. One of the differences between our method and SimCLR method is that our pipeline does not use a projection head during the learning stage. The performance of our method (without using a projection head) and SimCLR method (using a projection head) is listed in Table 31. Our method outperforms SimCLR on three different kinship datasets. Qualitative results are shown in Figure 50.

## 6.5 CONCLUSION

This is the first study on video-based transformers for the kinship verification task in a semi-supervised manner, *i.e.* pre-training on face datasets without kinship labels/relations. To this end, a kinship-oriented augmentation method, Video-kin augmentation, is proposed to enable the model to learn kinship-like distributions based on the pre-training on face video datasets. Large scale experiments are conducted and show that our proposed framework achieves state-of-the-art performance on the Nemo-kinship dataset.

---

## SUMMARY AND CONCLUSION

---

### 7.1 SUMMARY

The main purpose of this thesis is to analyze and study vision based kinship recognition in a real scenario. The thesis analyzes the current relevant research methods and explores the difficulties of the current application of kinship recognition in the real world. Based on these difficulties, the thesis proposes its basic methods in different chapters. The specific summary of each chapter is as follows:

#### ***Chapter 2: A Survey on Kinship Verification***

By reviewing the existing literature on kinship verification, we can better understand the challenges and successes in kinship recognition. Chapter 2 gives the answer to the first research question and presents a review of public datasets and representative methods for kinship verification. Representative methods are categorized and presented. To address the first research question ("What is kinship verification and what are the challenges"), this chapter studies current kinship challenges according to intrinsic factors (face *i.e.*, differences in facial appearance) and extrinsic factors (acquisition *i.e.*, varying imaging conditions). New promising directions are discussed based on current advances in kinship research. For instance, open-set kinship verification and debiasing kinship verification are largely ignored so far. They are promising for the kinship verification task in the future. The review notes that there is a need for more kinship datasets, particularly video-based ones, and introduces a new video dataset as a benchmark for child-adult kinship verification. This dataset consists of 248 subjects from 85 families. This benchmark is used to systematically test and analyze current state-of-the-art methods. Based on Chapter 2, several works targeting exploring kinship recognition in the real world are presented in the following chapters.

#### ***Chapter 3: Kinship Identification through Joint Learning***

Chapter 3 addresses the second research question ("How can kin types be better verified when facing unbalanced distribution in real scenarios?") by presenting a new method for kinship identification through joint learning. A training procedure on mixed-dataset is proposed. The unbalanced training data type between non-kin and other kin types make the model learn better discriminative feature among different kin types. Experimental results show that joint learning with kinship verification and identification improves

the performance of kinship identification. Chapter 3 proposes a basic approach on kinship identification in the real scenario. Since this method is not restricted to any neural network, a better architecture can further improve the performance in kinship identification.

#### ***Chapter 4: Identity Invariant Age Transfer for Kinship Verification of Child-Adult Images***

Chapter 2 introduces challenges of kinship verification. In Chapter 4, the issue of aging in kinship verification is discussed. Aging can impact the performance of kinship verification models in different ways. For instance, the images of a person of different ages can influence the performance of kinship verification. The image pairs of an older person with an adult can also affect the performance of the kinship verification models. Specifically, this chapter focuses on a more specific and overlooked situation, tackling kinship verification on child-adult images. To address this task, we propose a novel Identity-Invariance-Aging-Transferring approach that extracts identity-invariant information while removing the effects of aging as much as possible. In this way, the identity-invariant feature of each sample is extracted and transferred into a similar age distribution. Moreover, a kinship mapping module is used to compute the improved kinship-related information from the features of the CAT Module. The results show that, compared to the handcrafted feature, the transferred features capture the hidden features of genetic relationships and provide more robust results for child-related pairs.

#### ***Chapter 5: Kinship Similarity for Open Sets***

When addressing the fourth research question ("How can we improve kinship recognition when facing unknown classes?"), it is essential to identify and understand the different types of unknown classes that can exist in real-world scenarios. As noted in Chapter 5, unknown pairs may exist without any genetic relationship or with unlabeled kin relationships in open-set collections. Unknown pairs without genetic relationships can not be clearly clarified into one specific relation type. These unknown pairs without genetic relationships are taken as negative pairs. In reality, there are also some far kin-relationship but unlabeled pairs. These types may lead to (hard) negative samples, possibly affecting the performance of close-set trained models in a negative way. These types of samples require special attention and consideration when developing kinship recognition methods.

To answer the fourth research question, Chapter 5 proposes a new subtask of kinship recognition to determine kinship similarity in open sets. A method is proposed to determine family relationships and their corresponding degrees of kinship. The proposed method is pairwise-based and uses mutual information from positive pairs in a hierarchical way. Experiments and an ablation study show that our method outperforms the compared methods. Our model is able to separate kinship categories properly and generates uniform similarity distributions.

By proposing such Open-set Kinship Similarity Measurement, we hope there will be more approaches in the future. There are several potential benefits. Firstly, a good method on OKSM can utilize kinship similarities to identify (hard) non-kin samples

among potential kin pairs, which can help create more accurate collections of close-set kin samples for kinship-related tasks. Additionally, the kin-related pairs without labels (*i.e.* kinship degree) in the public dataset can be picked. Such unlabeled but kin-related samples can be used to enhance and enlarge kinship datasets.

### ***Chapter 6: Kinship Verification in Videos using Semi-Supervised Learning***

Chapter 6 gives an answer to the fifth research question ("How can we explore off-the-shelf knowledge from pretrained facial networks for kinship verification with limited kinship dataset?"). To better utilize the off-the-shelf knowledge from pretrained facial networks, we give a study on video-based transformers for the kinship verification task in a semi-supervised manner. A kinship-oriented augmentation method, Video-kin augmentation, is proposed to enable the model to learn kinship-like distributions based during the pretraining on facial video datasets. Large-scale experiments show that our proposed framework achieves state-of-the-art performance on the Nemo-kinship dataset.

## 7.2 CONCLUSION

The main contributions of this thesis can be divided into the following four points: First, this thesis comprehensively analyzes and summarizes the related work and datasets for kinship recognition. Second, this thesis proposes a new dataset for kinship prediction. Third, this thesis explores some possible challenges in the real world scenario and proposes the basic methods respectively. Fourth, this thesis shares the relevant codes and proposes some promising directions.





---

## APPENDIX

---

### A.1 SOFTWARE & REPOSITORIES

Overall, the authors provide relevant codes and GitHub repositories for each chapter:

- The public code for our kinship verification survey in Chapter 2 is provided at [https://github.com/we-wan/kin\\_sv](https://github.com/we-wan/kin_sv).
- The public code for our JLNet in Chapter 3 is provided at <https://github.com/we-wan/JLNet>.
- The public code for our CATNet in Chapter 4 is provided at <https://github.com/anonymous-sdfasdfa/-catnet->.
- The public code for our OKSM method in Chapter 5 is provided at <https://github.com/anonymous-fdfdklkl/OKSM>.
- The public code for our Kinship-transformer (KT) in Chapter 6 is provided at <https://github.com/kdafsdnmdfa/-1848->.





---

## BIBLIOGRAPHY

---

- [1] Y. Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:721–732, 1997.
- [2] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481. Springer, 2004.
- [3] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:2037–2041, 2006.
- [4] P. Alirezazadeh, A. Fathi, and F. Abdali-Mohammadi. A genetic algorithm-based feature selection for kinship verification. *IEEE Signal Processing Letters*, 22(12):2459–2463, 2015.
- [5] P. Alirezazadeh, A. Fathi, and F. Abdali-Mohammadi. A genetic algorithm-based feature selection for kinship verification. *IEEE Signal Processing Letters*, 22(12):2459–2463, 2015.
- [6] M. Almuashi, S. Z. M. Hashim, D. Mohamad, M. H. Alkawaz, and A. Ali. Automated kinship verification and identification through human facial images: a survey. *Multimedia Tools and Applications*, 76(1):265–307, 2017.
- [7] M. Almuashi, S. Z. M. Hashim, N. Yusoff, K. N. Syazwan, and F. Ghabban. Siamese convolutional neural network and fusion of the best overlapping blocks for kinship verification. *Multimedia Tools and Applications*, pages 1–32, 2022.
- [8] A. Alvergne, C. Faurie, and M. Raymond. Differential facial resemblance of young children to their parents: who do children look like more? *Evolution and Human Behavior*, 28(2):135–144, 2007.
- [9] A. Alvergne, R. Oda, C. Faurie, A. Matsumoto-Oda, V. Durand, and M. Raymond. Cross-cultural perceptions of facial resemblance between kin. *Journal of Vision*, 9(6):23–23, 2009.
- [10] A. Ariyaeinia, C. Morrison, A. Malegaonkar, and S. Black. A test of the effectiveness of speaker verification for differentiating between identical twins. *Science and Justice*, 48(4):182–186, 2008.
- [11] W. Bank. *Technology Landscape for Digital Identification*. World Bank, 2018.
- [12] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Recognition using class specific linear projection, 1997.
- [13] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
- [14] A. Bendale and T. E. Boulton. Towards open set deep networks. In *Computer Vision and Pattern Recognition*, pages 1563–1572. IEEE, 2016.
- [15] M. Bessaoudi, A. Ouamane, M. Belahcene, A. Chouchane, E. Boutellaa, and S. Bourennane. Multilinear side-information based discriminant analysis for face and kinship verification in the wild. *Neurocomputing*, 329:267–278, 2019.
- [16] A. R. Blaustein and R. K. O’Hara. Genetic control for sibling recognition? *Nature*, 290(5803):246–248, 1981.
- [17] A. Bottinok, I. U. Islam, and T. F. Vieira. A multi-perspective holistic approach to kinship verification in the wild. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 2, pages 1–6. IEEE, 2015.

- [18] E. Boutellaa, M. B. López, S. Ait-Aoudia, X. Feng, and A. Hadid. Kinship verification from videos using spatio-temporal texture features and deep learning. *arXiv preprint arXiv:1708.04069*, 2017.
- [19] E. Boutellaa, M. B. López, S. Ait-Aoudia, X. Feng, and A. Hadid. Kinship verification from videos using spatio-temporal texture features and deep learning. *arXiv preprint arXiv:1708.04069*, 2017.
- [20] R. L. Burch and G. G. Gallup Jr. Perceptions of paternal resemblance predict family violence. *Evolution and Human Behavior*, 21(6):429–435, 2000.
- [21] X. Cai, C. Wang, B. Xiao, X. Chen, and J. Zhou. Deep nonlinear metric learning with independent subspace analysis for face verification. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 749–752. ACM, 2012.
- [22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
- [23] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.
- [24] D. Carpenter, A. McLeod, C. Hicks, and M. Maasberg. Privacy and biometrics: An empirical examination of employee concerns. *Information Systems Frontiers*, 20(1):91–110, 2018.
- [25] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision*, pages 768–783. Springer, 2014.
- [26] G. Chen, P. Peng, X. Wang, and Y. Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.
- [27] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, and Y. Tian. Learning open set network with discriminative reciprocal points. In *European Conference on Computer Vision*, pages 507–522. Springer, 2020.
- [28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.
- [29] X. Chen, L. An, S. Yang, and W. Wu. Kinship verification in multi-linear coherent spaces. *Multimedia Tools and Applications*, 76(3):4105–4122, 2017.
- [30] X. Chen, C. Li, X. Zhu, L. Zheng, Y. Chen, S. Zheng, and C. Yuan. Deep discriminant generation-shared feature learning for image-based kinship verification. *Signal Processing: Image Communication*, page 116543, 2021.
- [31] X. Chen, C. Li, X. Zhu, L. Zheng, Y. Chen, S. Zheng, and C. Yuan. Deep discriminant generation-shared feature learning for image-based kinship verification. *SPIC*, 101:116543, 2022.
- [32] Y.-Y. Chen, W. H. Hsu, and H.-Y. M. Liao. Discovering informative social subgraphs and predicting pairwise relationships from group photos. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 669–678, 2012.
- [33] A. Chergui, S. Ouchtati, S. Mavromatis, S. E. Bekhouche, M. Lashab, and J. Sequeira. Kinship verification through facial images using cnn-based features. *Traitement du Signal*, 37(1), 2020.
- [34] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021.
- [35] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

- [36] S. Chung and N. Blake. Family reunification after disasters. *Clinical Pediatric Emergency Medicine*, 15(4):334–342, 2014.
- [37] E. Dahan and Y. Keller. Selfkin: self adjusted deep model for kinship verification. *arXiv preprint arXiv:1809.08493*, 2018.
- [38] M. F. Dal Martello and L. T. Maloney. Where are kin recognition signals in the human face? *Journal of Vision*, 6(12):2–2, 2006.
- [39] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. Ieee, 2005.
- [40] M. Daly and M. I. Wilson. Whom are newborn babies said to resemble? *Ethology and Sociobiology*, 3(2):69–78, 1982.
- [41] A. R. Dandekar and M. Nimberte. A survey: Verification of family relationship from parents and child facial images. In *2014 IEEE Students' Conference on Electrical, Electronics and Computer Science*, pages 1–6. IEEE, 2014.
- [42] L. M. DeBruine. Facial resemblance enhances trust. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1498):1307–1312, 2002.
- [43] L. M. DeBruine, F. G. Smith, B. C. Jones, S. C. Roberts, M. Petrie, and T. D. Spector. Kin recognition signals in adult faces. *Vision Research*, 49(1):38–43, 2009.
- [44] A. Dehghan, E. G. Ortiz, R. Villegas, and M. 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.
- [45] M. M. Dehshibi and J. Shanbehzadeh. Cubic norm and kernel-based bi-directional pca: toward age-aware facial kinship verification. *The Visual Computer*, 35(1):23–40, 2019.
- [46] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.
- [47] J. Deng, J. Guo, N. Xue, and S. 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.
- [48] H. 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.
- [49] H. Dibeklioglu, A. Ali Salah, and T. Gevers. Like father, like son: Facial expression dynamics for kinship verification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1497–1504, 2013.
- [50] H. Dibeklioglu, A. Ali Salah, and T. Gevers. Like father, like son: Facial expression dynamics for kinship verification. In *International Conference on Computer Vision*, pages 1497–1504, 2013.
- [51] H. Dibeklioglu, A. A. Salah, and T. Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*, pages 525–538. Springer, 2012.
- [52] Z. Ding, S. Suh, J.-J. Han, C. Choi, and Y. Fu. Discriminative low-rank metric learning for face recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6. IEEE, 2015.
- [53] G.-N. Dong, C.-M. Pun, and Z. Zhang. Deep collaborative multi-modal learning for unsupervised kinship estimation. *IEEE Transactions on Information Forensics and Security*, 16:4197–4210, 2021.

- [54] F. Dornaika, I. Arganda-Carreras, and O. Serradilla. Transfer learning and feature fusion for kinship verification. *Neural Computing and Applications*, 32(11):7139–7151, 2020.
- [55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [56] Q. Duan, L. Zhang, and W. 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.
- [57] X. Duan and Z.-H. Tan. A feature subtraction method for image based kinship verification under uncontrolled environments. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1573–1577. IEEE, 2015.
- [58] J. Dubisch. Gender, kinship, and religion: ‘reconstructing’ the anthropology of greece. *Contested identities: Gender and kinship in modern Greece*, pages 29–46, 1991.
- [59] I. C. Duta, L. Liu, F. Zhu, and L. Shao. Improved residual networks for image and video recognition. In *International Conference on Pattern Recognition*, pages 9415–9422. IEEE, 2021.
- [60] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers. In *International Conference on Computer Vision*, pages 6824–6835, 2021.
- [61] R. Fang, A. C. Gallagher, T. Chen, and A. Loui. Kinship classification by modeling facial feature heredity. In *2013 IEEE International Conference on Image Processing*, pages 2983–2987. IEEE, 2013.
- [62] R. Fang, K. D. Tang, N. Snavely, and T. Chen. Towards computational models of kinship verification. In *2010 IEEE International Conference on Image Processing*, pages 1577–1580. IEEE, 2010.
- [63] Y. Fang, Y. Yan, S. Chen, H. Wang, and C. Shu. Sparse similarity metric learning for kinship verification. In *2016 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2016.
- [64] D. M. Fessler and C. D. Navarrete. Third-party attitudes toward sibling incest: Evidence for westermarck’s hypotheses. *Evolution and Human Behavior*, 25(5):277–294, 2004.
- [65] B. Gao and L. Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.
- [66] P. Gao, S. Xia, J. Robinson, J. Zhang, C. Xia, M. Shao, and Y. Fu. What will your child look like? dna-net: Age and gender aware kin face synthesizer. *arXiv preprint arXiv:1911.07014*, 2019.
- [67] C. Garvie and J. Frankle. Facial-recognition software might have a racial bias problem. *The Atlantic*, 7, 2016.
- [68] Z. Ge, S. Demnyanov, Z. Chen, and R. Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017.
- [69] C. Geng, S.-j. Huang, and S. Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 3614–3631, 2020.
- [70] M. Georgopoulos, Y. Panagakis, and M. Pantic. Modeling of facial aging and kinship: A survey. *Image and Vision Computing*, 80:58–79, 2018.
- [71] J. Giordano, M. O’Reilly, H. Taylor, and N. Dogra. Confidentiality and autonomy: The challenge (s) of offering research participants a choice of disclosing their identity. *Qualitative Health Research*, 17(2):264–275, 2007.
- [72] A. Goel, A. Singh, A. Agarwal, M. Vatsa, and R. Singh. Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2018.

- [73] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.
- [74] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [75] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [76] A. Goyal, T. Meenpal, and M. Mukherjee. Family classification and kinship verification from facial images in the wild. *Machine Vision and Applications*, 33(6):1–18, 2022.
- [77] M. Gunther, S. Cruz, E. M. Rudd, and T. E. Boulton. Toward open-set face recognition. In *Computer Vision and Pattern Recognition Workshop*, pages 71–80. IEEE, 2017.
- [78] Y. Guo, H. Dibeklioglu, and L. Van der Maaten. Graph-based kinship recognition. In *2014 22nd International Conference on Pattern Recognition*, pages 4287–4292. IEEE, 2014.
- [79] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016.
- [80] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006.
- [81] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.
- [82] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [83] S. Hörmann, M. Knoche, and G. 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.
- [84] J. Hu, J. Lu, and Y.-P. Tan. Sharable and individual multi-view metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(9):2281–2288, 2017.
- [85] J. Hu, J. Lu, Y.-P. Tan, J. Yuan, and J. Zhou. Local large-margin multi-metric learning for face and kinship verification. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(8):1875–1891, 2017.
- [86] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *Asian Conference on Computer Vision*, pages 252–267. Springer, 2014.
- [87] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.
- [88] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2518–2525. IEEE, 2012.
- [89] S. Huang, J. Lin, L. Huangfu, Y. Xing, J. Hu, and D. D. Zeng. Adaptively weighted k-tuple metric network for kinship verification. *IEEE Transactions on Cybernetics*, 2022.
- [90] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.
- [91] H. Jin, Q. Liu, H. Lu, and X. Tong. Face detection using improved lbp under bayesian framework. In *Third International Conference on Image and Graphics (ICIG'04)*, pages 306–309. IEEE, 2004.
- [92] I. Jo, J. Kim, H. Kang, Y.-D. Kim, and S. Choi. Open set recognition by regularising classifier with fake data generated by generative adversarial networks. In *IEEE international conference on acoustics, speech and signal processing*, pages 2686–2690. IEEE, 2018.

- [93] G. Kaminski, S. Dridi, C. Graff, and E. Gentaz. Human ability to detect kinship in strangers' faces: effects of the degree of relatedness. *Proceedings of the Royal Society B: Biological Sciences*, 276(1670):3193–3200, 2009.
- [94] J. E. Kelly. Kinship and religious politics among catholic families in england, 1570–1640. *History*, 94(315):328–343, 2009.
- [95] M. J. Khoury, B. H. Cohen, E. L. Diamond, G. A. Chase, and V. A. McKusick. Inbreeding and prereproductive mortality in the old order amish. i. genealogic epidemiology of inbreeding. *American Journal of Epidemiology*, 125 3:453–61, 1987.
- [96] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann. Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146, 2020.
- [97] J. C. Klontz and A. K. Jain. A case study of automated face recognition: The boston marathon bombings suspects. *Computer*, 46(11):91–94, 2013.
- [98] N. Kohli, R. Singh, and M. Vatsa. Self-similarity representation of weber faces for kinship classification. In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 245–250. IEEE, 2012.
- [99] N. Kohli, M. Vatsa, R. Singh, A. Noore, and A. Majumdar. Hierarchical representation learning for kinship verification. *IEEE Transactions on Image Processing*, 26(1):289–302, 2016.
- [100] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore. Supervised mixed norm autoencoder for kinship verification in unconstrained videos. *IEEE Transactions on Image Processing*, 28(3):1329–1341, 2018.
- [101] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore. Supervised mixed norm autoencoder for kinship verification in unconstrained videos. *IEEE Transactions on Image Processing*, 28(3):1329–1341, 2018.
- [102] P. Korshunov and S. Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.
- [103] L. Kou, X. Zhou, M. Xu, and Y. Shang. Learning a genetic measure for kinship verification using facial images. *Mathematical Problems in Engineering*, 2015, 2015.
- [104] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.
- [105] H. J. Künzel. Automatic speaker recognition of identical twins. *International Journal of Speech, Language and the Law*, 17(2), 2010.
- [106] O. Laiadi, A. Ouamane, A. Benakcha, A. Taleb-Ahmed, and A. Hadid. Kinship verification based deep and tensor features through extreme learning machine. In *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, pages 1–4. IEEE, 2019.
- [107] O. Laiadi, A. Ouamane, A. Benakcha, A. Taleb-Ahmed, and A. Hadid. Tensor cross-view quadratic discriminant analysis for kinship verification in the wild. *Neurocomputing*, 377:286–300, 2020.
- [108] D. Lelis and D. Borges. Facial kinship verification with large age variation using deep linear metric learning. *Journal of Image and Graphics*, 7, 07 2019.
- [109] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3499–3506, 2013.
- [110] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2659, 2017.

- [111] L. Li, X. Feng, X. Wu, Z. Xia, and A. Hadid. Kinship verification from faces via similarity metric based convolutional neural network. In *International Conference on Image Analysis and Recognition*, pages 539–548. Springer, 2016.
- [112] W. Li, J. Lu, A. Wuerkaixi, J. Feng, and J. Zhou. Reasoning graph networks for kinship verification: from star-shaped to hierarchical. *IEEE Transactions on image processing*, 30:4947–4961, 2021.
- [113] W. Li, S. Wang, J. Lu, J. Feng, and J. Zhou. Meta-mining discriminative samples for kinship verification. *arXiv preprint arXiv:2103.15108*, 2021.
- [114] W. Li, Y. Zhang, K. Lv, J. Lu, J. Feng, and J. Zhou. Graph-based kinship reasoning network. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.
- [115] Z. Li, U. Park, and A. K. Jain. A discriminative model for age invariant face recognition. *IEEE Transactions on Information Forensics and Security*, 6(3):1028–1037, 2011.
- [116] J. Liang, J. Guo, S. Lao, and J. Li. Using deep relational features to verify kinship. In *CCF Chinese Conference on Computer Vision*, pages 563–573. Springer, 2017.
- [117] J. Liang, Q. Hu, C. Dang, and W. Zuo. Weighted graph embedding-based metric learning for kinship verification. *IEEE Transactions on Image Processing*, 28(3):1149–1162, 2018.
- [118] D. Lieberman, J. Tooby, and L. Cosmides. The architecture of human kin detection. *Nature*, 445(7129):727–731, 2007.
- [119] F. Liu, Z. Li, W. Yang, and F. Xu. Age-invariant adversarial feature learning for kinship verification. *Mathematics*, 10(3):480, 2022.
- [120] Q. Liu, A. Puthenputhussery, and C. Liu. Inheritable fisher vector feature for kinship verification. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE, 2015.
- [121] Q. Liu, A. Puthenputhussery, and C. Liu. A novel inheritable color space with application to kinship verification. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.
- [122] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, pages 10012–10022, 2021.
- [123] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.
- [124] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In *Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.
- [125] M. B. Lopez, A. Hadid, E. Boutellaa, J. Goncalves, V. Kostakos, and S. Hosio. Kinship verification from facial images and videos: human versus machine. *Machine Vision and Applications*, 29(5):873–890, 2018.
- [126] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [127] J. Lu, J. Hu, V. E. Liong, X. Zhou, A. Bottino, I. U. Islam, T. F. Vieira, X. Qin, X. Tan, S. Chen, et al. The fg 2015 kinship verification in the wild evaluation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–7. IEEE, 2015.
- [128] J. Lu, J. Hu, and Y.-P. Tan. Discriminative deep metric learning for face and kinship verification. *IEEE Transactions on Image Processing*, 26(9):4269–4282, 2017.

- [129] J. Lu, J. Hu, X. Zhou, Y. Shang, Y.-P. Tan, and G. Wang. Neighborhood repulsed metric learning for kinship verification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2594–2601. IEEE, 2012.
- [130] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):331–345, 2013.
- [131] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):331–345, 2013.
- [132] C. Luo, L. Jin, and J. Chen. Siman: Exploring self-supervised representation learning of scene text via similarity-aware normalization. In *Computer Vision and Pattern Recognition*, pages 1039–1048, 2022.
- [133] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205. IEEE, 1998.
- [134] S. Mahpod and Y. Keller. Kinship verification using multiview hybrid distance learning. *Computer Vision and Image Understanding*, 167:28–36, 2018.
- [135] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *International Conference on Computer Vision*, pages 2794–2802, 2017.
- [136] I. Masi, Y. Wu, T. Hassner, and P. Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 471–478. IEEE, 2018.
- [137] D. Miller, N. Sunderhauf, M. Milford, and F. Dayoub. Class anchor clustering: A loss for distance-based open set recognition. In *Winter Conference on Applications of Computer Vision*, pages 3570–3578, 2021.
- [138] U. B. Mir, A. K. Kar, and M. P. Gupta. Digital identity evaluation framework for social welfare. In *International Working Conference on Transfer and Diffusion of IT*, pages 401–414. Springer, 2020.
- [139] M. Mukherjee, T. Meenpal, and A. Goyal. Fusekin: Weighted image fusion based kinship verification under unconstrained age group. *Journal of Visual Communication and Image Representation*, 84:103470, 2022.
- [140] M. Murphy. Variations in kinship networks across geographic and social space. *Population and Development Review*, 34(1):19–49, 2008.
- [141] S. K. Nadimpalli, S. Prasad, and P. Raghava. Kinship terms in telugu and english. *International Journal of Humanities and Social Science Invention*, 3(4):44–46, 2014.
- [142] A. Nandy and S. S. Mondal. Kinship verification using deep siamese convolutional neural network. In *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019.
- [143] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li. Open set learning with counterfactual images. In *European Conference on Computer Vision*, pages 613–628, 2018.
- [144] D. Neimark, O. Bar, M. Zohar, and D. Asselmann. Video transformer network. In *International Conference on Computer Vision*, pages 3163–3172, 2021.
- [145] T.-D. H. Nguyen, H.-N. H. Nguyen, and H. Dao. Recognizing families through images with pretrained encoder. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 887–891. IEEE, 2020.



- [146] C. Ober, T. Hyslop, and W. W. Hauck. Inbreeding effects on fertility in humans: evidence for reproductive compensation. *The American Journal of Human Genetics*, 64(1):225–231, 1999.
- [147] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 582–585. IEEE, 1994.
- [148] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [149] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.
- [150] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.
- [151] B. Patel, R. Maheshwari, and B. Raman. Evaluation of periocular features for kinship verification in the wild. *Computer Vision and Image Understanding*, 160:24–35, 2017.
- [152] G. Peleg, G. Katzir, O. Peleg, M. Kamara, L. Brodsky, H. Hel-Or, D. Keren, and E. Nevo. Hereditary family signature of facial expression. *Proceedings of the National Academy of Sciences*, 103(43):15921–15926, 2006.
- [153] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.
- [154] S. M. Platek and J. W. Thomson. Facial resemblance exaggerates sex-specific jealousy-based decisions. *Evolutionary Psychology*, 5(1):147470490700500113, 2007.
- [155] A. Puthenputhussery, Q. Liu, and C. Liu. Sift flow based genetic fisher vector feature for kinship verification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2921–2925. IEEE, 2016.
- [156] X. Qin, D. Liu, and D. Wang. A literature survey on kinship verification through facial images. *Neurocomputing*, 377:213–224, 2020.
- [157] X. Qin, D. Liu, and D. Wang. A novel factor analysis-based metric learning method for kinship verification. *Multimedia Tools and Applications*, 81(8):11049–11070, 2022.
- [158] X. Qin, X. Tan, and S. Chen. Tri-subject kinship verification: Understanding the core of a family. *IEEE Transactions on Multimedia*, 17(10):1855–1867, 2015.
- [159] X. Qin, X. Tan, and S. Chen. Mixed bi-subject kinship verification via multi-view multi-task learning. *Neurocomputing*, 214:350–357, 2016.
- [160] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872, 2018.
- [161] N. Ramanathan and R. Chellappa. Modeling age progression in young faces. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 387–394. IEEE, 2006.
- [162] Ramanathan, Narayanan and Chellappa, Rama. Modeling shape and textural variations in aging faces. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8. IEEE, 2008.
- [163] J. P. Robinson, Z. Khan, Y. Yin, M. Shao, and Y. Fu. Families in wild multimedia: A multimodal database for recognizing kinship. *arXiv preprint arXiv:2007.14509*, 2020.
- [164] J. P. Robinson, C. Qin, M. Shao, M. A. Turk, R. Chellappa, and Y. Fu. The 5th recognizing families in the wild data challenge: Predicting kinship from faces. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–07. IEEE, 2021.

- [165] J. P. Robinson, M. Shao, and Y. 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.
- [166] J. P. Robinson, M. Shao, Y. Wu, and Y. Fu. Families in the wild (fiw): Large-scale kinship image database and benchmarks. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 242–246. ACM, 2016.
- [167] J. P. Robinson, M. Shao, Y. Wu, and Y. Fu. Family in the wild (FIW): A large-scale kinship recognition database. *CoRR*, abs/1604.02182, 2016.
- [168] J. P. Robinson, M. Shao, Y. Wu, H. Liu, T. Gillis, and Y. Fu. Visual kinship recognition of families in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2624–2637, 2018.
- [169] J. P. Robinson, M. Shao, Y. Wu, H. Liu, T. Gillis, and Y. Fu. Visual kinship recognition of families in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2624–2637, 2018.
- [170] J. P. Robinson, M. Shao, H. Zhao, Y. Wu, T. Gillis, and Y. Fu. 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. ACM, 2017.
- [171] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2012.
- [172] W. J. Scheirer, L. P. Jain, and T. E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2317–2324, 2014.
- [173] E. Selinger and W. Hartzog. The inconstentability of facial surveillance. *Loy. L. Rev.*, 66:33, 2020.
- [174] I. Serroui, O. Laiadi, A. Ouamane, F. Dornaika, and A. Taleb-Ahmed. Knowledge-based tensor subspace analysis system for kinship verification. *Neural Networks*, 151:222–237, 2022.
- [175] A. 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.
- [176] M. Shao, S. Xia, and Y. Fu. Genealogical face recognition based on ub kinface database. In *Computer Vision and Pattern Recognition 2011 Workshops*, pages 60–65. IEEE, 2011.
- [177] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM Sigsac Conference on Computer and Communications Security*, pages 1528–1540, 2016.
- [178] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3):1–30, 2019.
- [179] G. Somanath and C. Kambhamettu. Can faces verify blood-relations? In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 105–112. IEEE, 2012.
- [180] C. Song and H. Yan. Kinmix: A data augmentation approach for kinship verification. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.
- [181] J. Song, J. Zhang, L. Gao, X. Liu, and H. T. Shen. Dual conditional gans for face aging and rejuvenation. In *International Joint Conference on Artificial Intelligence*, pages 899–905, 2018.
- [182] Z. Stone, T. Zickler, and T. Darrell. Toward large-scale face recognition using social network context. *Proceedings of the IEEE*, 98(8):1408–1415, 2010.

- [183] Q. Sun, B. Schiele, and M. Fritz. A domain based approach to social relation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3481–3490, 2017.
- [184] Y. Sun, X. Wang, and X. Tang. 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.
- [185] A. Surmiak et al. Confidentiality in qualitative research involving vulnerable participants: Researchers’ perspectives. In *Forum: Qualitative Social Research*, volume 19, pages 393–418. Freie Universität Berlin, 2018.
- [186] J. H. Sweet. *Recreating Africa: culture, kinship, and religion in the African-Portuguese world, 1441-1770*. Univ of North Carolina Press, 2003.
- [187] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.
- [188] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591. IEEE, 1991.
- [189] A. Van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.
- [190] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103. ACM, 2008.
- [191] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pages 1096–1103, 2008.
- [192] M. Wang and W. Deng. Deep face recognition: A survey. *CoRR*, abs/1804.06655, 2018.
- [193] M. Wang and W. Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.
- [194] M. Wang, J. Feng, X. Shu, Z. Jie, and J. Tang. Photo to family tree: Deep kinship understanding for nuclear family photos. In *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, pages 41–46, 2018.
- [195] M. Wang, X. Shu, J. Feng, X. Wang, and J. Tang. Deep multi-person kinship matching and recognition for family photos. *Pattern Recognition*, 105:107342, 2020.
- [196] M. Wang, Zechao Li, Xiangbo Shu, Jingdong, and J. Tang. Deep kinship verification. In *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2015.
- [197] S. Wang, Z. Ding, and Y. Fu. Cross-generation kinship verification with sparse discriminative metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2783–2790, 2018.
- [198] S. Wang, J. P. Robinson, and Y. Fu. Kinship verification on families in the wild with marginalized denoising metric learning. In *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, pages 216–221. IEEE, 2017.
- [199] S. Wang and H. Yan. Discriminative sampling via deep reinforcement learning for kinship verification. *Pattern Recognition Letters*, 138:38–43, 2020.
- [200] W. Wang, Z. Cui, Y. Yan, J. Feng, S. Yan, X. Shu, and N. Sebe. Recurrent face aging. In *Computer Vision and Pattern Recognition*, pages 2378–2386, 2016.

- [201] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *International Conference on Computer Vision*, pages 568–578, 2021.
- [202] W. Wang, S. You, and T. Gevers. Kinship identification through joint learning using kinship verification ensembles. In *European Conference on Computer Vision*, pages 613–628. Springer, 2020.
- [203] X. Wang, G. Guo, M. Merler, N. C. Codella, M. Rohith, J. R. Smith, and C. Kambhamettu. Leveraging multiple cues for recognizing family photos. *Image and Vision Computing*, 58:61–75, 2017.
- [204] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *2009 IEEE 12th International Conference on Computer Vision*, pages 32–39. IEEE, 2009.
- [205] X. Wang and C. Kambhamettu. Leveraging appearance and geometry for kinship verification. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5017–5021. IEEE, 2014.
- [206] Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, and T. Zhang. Orthogonal deep features decomposition for age-invariant face recognition. In *European Conference on Computer Vision*, pages 738–753, 2018.
- [207] Z. Wang, J. Chen, and J. Hu. Multi-view cosine similarity learning with application to face verification. *Mathematics*, 10(11):1800, 2022.
- [208] Z. Wang, X. Tang, W. Luo, and S. Gao. Face aging with identity-preserved conditional generative adversarial networks. In *Computer Vision and Pattern Recognition*, pages 7939–7947, 2018.
- [209] M. Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 2019.
- [210] C. Wissler. *The American Indian: An introduction to the anthropology of the New World*. Oxford University Press, 1922.
- [211] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition*, pages 529–534. IEEE, 2011.
- [212] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Real-Life Images workshop at the European Conference on Computer Vision*, October 2008.
- [213] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *European Conference on Computer Vision*, pages 3–19, 2018.
- [214] X. Wu, X. Feng, X. Cao, X. Xu, D. Hu, M. B. López, and L. Liu. Facial kinship verification: A comprehensive review and outlook. *International Journal of Computer Vision*, pages 1–32, 2022.
- [215] X. Wu, X. Feng, L. Li, E. Boutellaa, and A. Hadid. Kinship verification based on deep learning. In *Deep Learning in Object Detection and Recognition*, pages 113–132. Springer, 2019.
- [216] X. Wu, E. Granger, and X. Feng. Audio-visual kinship verification. *arXiv preprint arXiv:1906.10096*, 2019.
- [217] Y. Wu, Z. Ding, H. Liu, J. Robinson, and Y. Fu. Kinship classification through latent adaptive subspace. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pages 143–149. IEEE, 2018.
- [218] S. Xia, M. Shao, and Y. Fu. Kinship verification through transfer learning. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.
- [219] S. Xia, M. Shao, and Y. Fu. Kinship verification through transfer learning. In *Twenty-Second International Joint Conference on Artificial Intelligence*, page 2539–2544, 2011.

- [220] S. Xia, M. Shao, J. Luo, and Y. Fu. Understanding kin relationships in a photo. *IEEE Transactions on Multimedia*, 14(4):1046–1056, 2012.
- [221] S. Xia, M. Shao, J. Luo, and Y. Fu. Understanding kin relationships in a photo. *IEEE Transactions on Multimedia*, 14(4):1046–1056, 2012.
- [222] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang. Vision transformer with deformable attention. In *Computer Vision and Pattern Recognition*, pages 4794–4803, 2022.
- [223] M. Xu and Y. Shang. Kinship measurement on face images by structured similarity fusion. *IEEE Access*, 4:10280–10287, 2016.
- [224] H. Yan. Kinship verification using neighborhood repulsed correlation metric learning. *Image and Vision Computing*, 60:91–97, 2017.
- [225] H. Yan. Learning discriminative compact binary face descriptor for kinship verification. *Pattern Recognition Letters*, 117:146–152, 2019.
- [226] H. Yan and J. Hu. Video-based kinship verification using distance metric learning. *Pattern Recognition*, 75:15–24, 2018.
- [227] H. Yan and J. Lu. *Facial Kinship Verification: A Machine Learning Approach*. Springer, 2017.
- [228] H. Yan, J. Lu, W. Deng, and X. Zhou. Discriminative multimetric learning for kinship verification. *IEEE Transactions on Information Forensics and Security*, 9(7):1169–1178, 2014.
- [229] H. Yan, J. Lu, W. Deng, and X. Zhou. Discriminative multimetric learning for kinship verification. *IEEE Transactions on Information Forensics and Security*, 9(7):1169–1178, 2014.
- [230] H. Yan, J. Lu, and X. Zhou. Prototype-based discriminative feature learning for kinship verification. *IEEE Transactions on Cybernetics*, 45(11):2535–2545, 2014.
- [231] H. Yan and C. Song. Multi-scale deep relational reasoning for facial kinship verification. *Pattern Recognition*, 110:107541, 2021.
- [232] H. Yan and S. Wang. Learning part-aware attention networks for kinship verification. *Pattern Recognition Letters*, 128:169–175, 2019.
- [233] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.
- [234] W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.
- [235] J. Yu, M. Li, X. Hao, and G. 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.
- [236] L. A. Zebrowitz and J. M. Montepare. Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, 2(3):1497–1517, 2008.
- [237] H. Zhang and V. M. Patel. Sparse representation-based open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1690–1696, 2016.
- [238] H. Zhang, X. Wang, and C.-C. J. Kuo. Deep kinship verification via appearance-shape joint prediction and adaptation-based approach. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3856–3860. IEEE, 2019.
- [239] J. Zhang, S. Xia, H. Pan, and A. K. Qin. A genetics-motivated unsupervised model for tri-subject kinship verification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2916–2920. IEEE, 2016.
- [240] K. Zhang, Y. Huang, C. Song, H. Wu, and L. Wang. Kinship verification with deep convolutional neural networks. In *British Machine Vision Conference*, 2015.

- [241] L. Zhang, Q. Duan, D. Zhang, W. Jia, and X. Wang. Advkin: Adversarial convolutional network for kinship verification. *IEEE Transactions on Cybernetics*, 2020.
- [242] P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh. Predicting failures of vision systems. In *Computer Vision and Pattern Recognition*, pages 3566–3573. IEEE, 2014.
- [243] T. Zhang, D. Tao, and J. Yang. Discriminative locality alignment. In *European Conference on Computer Vision*, pages 725–738. Springer, 2008.
- [244] Z. Zhang, Y. Chen, and V. Saligrama. Group membership prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3916–3924, 2015.
- [245] J. Zhao, Y. Cheng, Y. Cheng, Y. Yang, F. Zhao, J. Li, H. Liu, S. Yan, and J. Feng. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9251–9258, 2019.
- [246] Y. Zhao, L.-M. Po, X. Wang, Q. Yan, W. Shen, Y. Zhang, W. Liu, C.-K. Wong, C.-S. Pang, W. Ou, et al. Childpredictor: A child face prediction framework with disentangled learning. *IEEE Transactions on Multimedia*, 2022.
- [247] Y. Zhong and W. Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2020.
- [248] J. Zhou, F. Chen, A. Berry, M. Reed, S. Zhang, and S. Savage. A survey on ethical principles of ai and implementations. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 3010–3017. IEEE, 2020.
- [249] X. Zhou, J. Hu, J. Lu, Y. Shang, and Y. Guan. Kinship verification from facial images under uncontrolled conditions. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 953–956. ACM, 2011.
- [250] X. Zhou, K. Jin, M. Xu, and G. Guo. Learning deep compact similarity metric for kinship verification from face images. *Information Fusion*, 48:84–94, 2019.
- [251] X. Zhou, J. Lu, J. Hu, and Y. 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. ACM, 2012.
- [252] X. Zhou, Y. Shang, H. Yan, and G. Guo. Ensemble similarity learning for kinship verification from facial images in the wild. *Information Fusion*, 32:40–48, 2016.
- [253] X. Zhou, H. Yan, and Y. Shang. Kinship verification from facial images by scalable similarity fusion. *Neurocomputing*, 197:136–142, 2016.
- [254] X. Zhu, C. Li, X. Chen, X. Zhang, and X.-Y. Jing. Distance and direction based deep discriminant metric learning for kinship verification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022.

---

## SAMENVATTING

---

### 7.2 SAMENVATTING

Het hoofddoel van dit proefschrift is het analyseren en bestuderen van op visie gebaseerde verwantschapsherkenning in een reëel scenario. Het proefschrift analyseert de huidige relevante onderzoeksmethoden en onderzoekt de moeilijkheden van de huidige toepassing van verwantschapsherkenning in de echte wereld. Op basis van deze moeilijkheden stelt het proefschrift in verschillende hoofdstukken zijn basismethoden voor. De specifieke samenvatting van elk hoofdstuk is als volgt:

#### ***Hoofdstuk 2: Een onderzoek naar verwantschapsverificatie***

Door de bestaande literatuur over verificatie van verwantschap te bestuderen, kunnen we de uitdagingen en successen bij de erkenning van verwantschap beter begrijpen. Hoofdstuk 2 geeft het antwoord op de eerste onderzoeksvraag en geeft een overzicht van openbare datasets en representatieve methoden voor verificatie van verwantschap. Representatieve methoden worden gecategoriseerd en gepresenteerd. Om de eerste onderzoeksvraag ("Wat is verwantschapsverificatie en wat zijn de uitdagingen") te beantwoorden, bestudeert dit hoofdstuk de huidige uitdagingen in verwantschapsverificatie volgens intrinsieke factoren (bijvoorbeeld het gezicht) en extrinsieke factoren (bijvoorbeeld verwerving van data met variërende beeldvormingsomstandigheden). Nieuwe veelbelovende richtingen worden besproken op basis van de huidige vorderingen in verwantschapsonderzoek. Zo worden open-set verwantschapsverificatie en debiasing verwantschapsverificatie tot nu toe grotendeels genegeerd. Deze richtingen zijn veelbelovend voor de verwantschapsverificatietask in de toekomst. In de review wordt opgemerkt dat er behoefte is aan meer verwantschapsdatasets, met name op video gebaseerde datasets, en er wordt een nieuwe videodataset geïntroduceerd als maatstaf voor verificatie van verwantschap tussen kinderen en volwassenen. Deze dataset bestaat uit 248 proefpersonen uit 85 families. De benchmark wordt gebruikt om de huidige state-of-the-art methoden systematisch te testen en te analyseren. In de volgende hoofdstukken worden verschillende werken gepresenteerd die gericht zijn op het onderzoeken van verwantschapsherkenning in de echte wereld gebaseerd op de inzichten in Hoofdstuk 2.

#### ***Hoofdstuk 3: Identificatie van verwantschap door gezamenlijk leren***

Hoofdstuk 3 gaat in op de tweede onderzoeksvraag ("Hoe kunnen verwantschapstypen beter worden geverifieerd wanneer ze worden geconfronteerd met een onevenwichtige data verdeling in reële scenario's?") door een nieuwe methode te presenteren voor verwantschapsidentificatie door middel van gezamenlijk leren. Er wordt een trainingsprocedure voor een gemengde dataset voorgesteld. Het onevenwicht tussen niet-verwante en andere verwante typen in de trainingsset zorgt ervoor dat het model een beter onderscheid leert maken tussen verschillende verwante typen. Experimentele resultaten laten zien dat gezamenlijk leren met verwantschapsverificatie en identificatie de prestaties van verwantschapsidentificatie verbetert. Hoofdstuk 3 stelt een basisbenadering voor van verwantschapsidentificatie in het echte scenario. Aangezien deze

methode niet beperkt is tot een neurale netwerk, kan een betere architectuur de prestaties bij verwantschapsidentificatie verder verbeteren.

#### ***Hoofdstuk 4: Identiteitsinvariante leeftijdsoverdracht voor verwantschapsverificatie van afbeeldingen van kinderen en volwassenen***

Hoofdstuk 2 introduceert uitdagingen van verificatie van verwantschap. In Hoofdstuk 4 wordt het probleem van ouder worden bij verwantschapsverificatie besproken. Veroudering kan de prestaties van verwantschapsverificatiemodellen op verschillende manieren beïnvloeden. De afbeeldingen van een persoon van verschillende leeftijden kunnen bijvoorbeeld de uitvoering van verwantschapsverificatie beïnvloeden. De beeldparen van een oudere persoon met een volwassene kunnen ook van invloed zijn op de prestaties van de verwantschapsverificatiemodellen. In het bijzonder richt dit hoofdstuk zich op een meer specifieke en over het hoofd geziene situatie, namelijk het aanpakken van verwantschapsverificatie op afbeeldingen van kinderen en volwassenen. Om deze taak aan te pakken, stellen we een nieuwe identiteitsinvariante en leeftijdsoverdragende benadering voor die identiteit-invariante informatie extraheert en tegelijkertijd de effecten van veroudering zoveel mogelijk verwijdert. Op deze manier wordt het identiteitsinvariante kenmerk van elke beeldpaar geëxtraheerd en overgebracht naar een vergelijkbare leeftijdsverdeling. Bovendien wordt een module voorgesteld die verwantschap voorspelt om de verbeterde verwantschapsgereleerde informatie uit de kenmerken van de CAT-module te berekenen. De resultaten laten zien dat de overgedragen kenmerken – in vergelijking met de handgemaakte kenmerken – de latente kenmerken van genetische relaties beter kunnen vastleggen en meer robuuste resultaten opleveren voor kindgerelateerde paren.

#### ***Hoofdstuk 5: Verwantschapsovereenkomst voor open sets***

Bij het beantwoorden van de vierde onderzoeksvraag ("Hoe kunnen we de herkenning van verwantschap verbeteren wanneer we geconfronteerd worden met onbekende klassen?"), is het essentieel om de verschillende soorten onbekende klassen te identificeren en te begrijpen die kunnen bestaan in scenario's in de echte wereld. Zoals opgemerkt in Hoofdstuk 5, kunnen onbekende paren voorkomen zonder enige genetische verwantschap of met niet-gelabelde verwantschapsrelaties in open verzamelingen. Onbekende paren zonder genetische relaties kunnen niet duidelijk worden opgehelderd in één specifiek relatietype. Deze onbekende paren zonder genetische verwantschap worden als negatieve paren beschouwd. In werkelijkheid zijn er ook enkele verre verwanten, maar niet-gelabelde paren. Deze typen kunnen leiden tot (harde) negatieve steekproeven, waardoor mogelijk de prestaties van close-set getrainde modellen op een negatieve manier worden beïnvloed. Dit soort voorbeelden vereist speciale aandacht en aandacht bij het ontwikkelen van methoden voor het herkennen van verwantschap.

Om de vierde onderzoeksvraag te beantwoorden, stelt Hoofdstuk 5 een nieuwe subtaak van verwantschapsherkenning voor om verwantschapsovereenkomst in open verzamelingen te bepalen. Er wordt een methode voorgesteld om familierelaties en hun overeenkomstige verwantschapsgraden te bepalen. De voorgestelde methode is paarsgewijs gebaseerd en gebruikt wederzijdse informatie van positieve paren op een hiërarchische manier. Experimenten en een ablatiestudie tonen aan dat onze methode beter presteert dan de vergeleken methoden. Ons model is in staat verwantschapscategorieën goed te scheiden en genereert uniforme gelijkenisverdelingen.

Door een dergelijke open-set verwantschapsgelijkenismeting (OKSM, Engels) voor te stellen, hopen we dat er in de toekomst meer benaderingen zullen zijn. Er zijn verschillende potentiële voordelen. Ten eerste kan een goede methode op OKSM verwantschapsovereenkomsten ge-



bruiken om (harde) niet-verwante steekproeven te identificeren tussen potentiële verwantenparen, wat kan helpen bij het creëren van nauwkeurigere verzamelingen van close-set verwantenparen voor verwantschapsgerelateerde taken. Bovendien kunnen de verwantenparen zonder labels (*d.w.z.* verwantschapsgraad) in de openbare dataset worden gekozen. Dergelijke niet-gelabelde maar verwante paren kunnen worden gebruikt om verwantschapsdatasets te verbeteren en uit te breiden.

### ***Hoofdstuk 6: Verificatie van verwantschap in video's met behulp van semi-supervised learning***

Hoofdstuk 6 geeft een antwoord op de vijfde onderzoeksvraag ("Hoe kunnen we eerdere kennis uit voorgetrainde gezichtsnetwerken verkennen voor verwantschapsverificatie met een beperkte verwantschapsdataset?"). Om beter gebruik te maken van de eerdere kennis van vooraf getrainde gezichtsnetwerken, bestuderen we op-video-gebaseerde transformatoren-netwerken voor de verwantschapsverificatietask op een semi-gesuperviseerde manier. Een op verwantschap georiënteerde augmentatiemethode, *video-kin augmentatie*, wordt voorgesteld om het model in staat te stellen verwantschap-achtige verdelingen te leren op basis van de pre-training op gezichtsvideodatasets. Grootschalige experimenten tonen aan dat ons voorgestelde raamwerk state-of-the-art prestaties levert op de Nemo-kinship dataset.

### 7.3 GEVOLGTREKKING

De belangrijkste bijdragen van dit proefschrift kunnen worden onderverdeeld in de volgende vier punten: Ten eerste analyseert en vat dit proefschrift uitgebreid het gerelateerde werk en de datasets voor verwantschapsherkenning samen. Ten tweede stelt dit proefschrift een nieuwe dataset voor het voorspellen van verwantschap. Ten derde onderzoekt dit proefschrift enkele mogelijke uitdagingen in het echte wereldscenario en stelt respectievelijk de basismethoden voor. Ten vierde deelt dit proefschrift de relevante codes en stelt het enkele veelbelovende richtingen voor.



---

## ACKNOWLEDGMENTS

---

"Life is not a race, but a journey to be savored each step of the way." This is a quote I read back in high school. I am deeply grateful to nature for allowing me the opportunity to be born in this vast, boundless universe, and within the endless river of time, to embark on my journey on this small, solitary planet. During my journey, I have been fortunate enough to choose the path of pursuing a PhD at the University of Amsterdam, where I have had the chance to meet so many kind and beautiful people. I still remember the day I arrived in the Netherlands; it was drizzling as I stepped off the plane. Before even reaching my dormitory, I made my first stop at the University of Amsterdam and met my professors and fellow group members. That moment marked the beginning of this incredible journey. Each day spent here fills my heart with gratitude and appreciation.

First and foremost, I would like to express my sincerest gratitude to my promoter, Theo Gevers, and his family. During my time at the University of Amsterdam, Theo has provided me with immense support and care, both academically and personally. His scholarly attitude, professional expertise, and approach to life have been profoundly influential and beneficial to me. He has not only been my academic mentor, but also a life mentor I will cherish throughout my life. I am deeply grateful for Theo's dedication and unwavering care for me. Over the past four years, there have been countless touching moments. Theo has encouraged me during my low times. I remember after our go-karting team-building event, he personally drove us home. I also recall him teaching me how to play "blackjack". Of course, I must express my gratitude for his trust, as Theo entrusted me with the care of his plants at home. I would also like to extend my thanks to Theo's family. I still remember the joy we shared playing soccer during our barbecue gatherings. I am deeply appreciative of the book Anja gifted me, which I truly cherish. I once promised Theo that I would send him greetings at least every Christmas in the future, and I hope I can live up to that commitment.

In addition, I would like to extend my heartfelt gratitude to my co-promoters, Shaodi You and Sezer Karaoglu. I am grateful for Shaodi's guidance and assistance in my daily life, teaching me how to think critically while conducting research. I appreciate the care and encouragement Shaodi has provided in my personal life, as he often generously treated us to meals. I am also grateful for his easel, the snacks he brought back from his travels, and his lifelong support. I am grateful to Sezer for his support and assistance during my PhD journey, and for the inspiration and suggestions he provided during our meetings. His smiles and uplifting spirit have always encouraged me. I would like to thank Sezer for preparing delicious food for us during our group barbecues and team-building events.

I would also like to express my heartfelt appreciation to my committee members: prof. dr. Marcel Worring, prof. dr. Albert Salah, prof. dr. ir. Peter H. N. de With, dr. Arnoud Visser, and dr. Efstratios Gavves. They are all individuals who are deeply admired. I am grateful for their agreement to be my committee members and for their support and encouragement. Their valuable feedback on the thesis has been an invaluable treasure. Thanks Prof. dr. P. van Emde Boas for being the chairman of my PhD committee.

I am extremely grateful to the teachers in our group, with whom I have spent countless memorable moments. I would like to express my thanks to: Arnoud, Dennis, Dimitrios, Leo, and Martin. Each interaction with them has been incredibly beneficial. I appreciate Arnoud's help and suggestions on my thesis, as well as the times you joined us for volleyball games. Thank you, Martin, for letting me experience the Hololens. I am grateful for talking with Leo and Dimitrios. And finally, thank you, Dennis, for your daily help and support.

I would like to express my gratitude to my colleagues in the group, with whom I have spent a memorable time (*listed alphabetically by the first capital letter*): Anil, Dimitrije, Hanan, Hoang-An, Jian, Li, Melis, Minh, Ozzy, Partha, Qi, Rick, Ronny, Ruihong, Vladimir, Wei, Weijie, Xiaoyan, Yahui, Yang, Yunlu, Zhiwei. We encouraged and exchanged ideas with each other. Special thanks to Yahui, Rick, and Partha for their meticulous care and assistance in my daily life. I appreciate Yahui's help and communication during this time. I am grateful to Rick and Partha for serving as my paranymphs. Let's not forget 'The Other Corner'. Let's not forget our "Gan Bei" drinking nights. I also want to thank my senior colleagues, Jian Han and Wei Zeng, for their unwavering care and support. They helped me a lot. I am grateful for the engaging conversations with Minh, Anil, Hoang-An, and Hanan. Thank you, Ruihong, for organizing events, sharing meals, and bringing joy. Thanks, Li, for the daily chats and shared meals. I am grateful to Weijie & Xiaoyan for hosting delightful hotpot gatherings and gaming sessions. Thanks to Xiaoyan & Pingping for taking me on scenic drives and sharing drinks together. Thanks, Qi and Yunlu, I enjoy the conversations and assistance. Thank you, Ozzy, for preparing delicious barbecue food. I truly treasure the time spent with each one of you.

Thank you to my dear friends (*listed alphabetically by the first capital letter*): Aozhu, Aritra, Chang, Cong, Congfeng, Dan, David, Di, Dongwei, Fan, Finde, Gaosheng, Haochen, Hongyun, Jia-Hong, Jiahuan, Jie, Jiayi, Jiayun, Jin, Jianghua, Jun, Kefan, Kaijie, Kexin, Lichun, Mengdi, Ming, Morris, Na, Nguyen, Pengwan, Pengyu, Pingping, Qi, Qingkang, Shaojie, Shihan, Shiqi, Shuai, Shuo, Teng, Tao, Wangduo, Wangyuan, Weijia, Wenxi, Wenzhe, Xiaoxiao, Xiaotian, Xiaowei, Xinghui, Xuemei, Xue, Xuelai, Yan, Yanhua, Yangjun, Yixian, Ying, Yingjun, Yongtuo, Yue, Yunhua, Yunlu, Zehao, Zenglin, Zeshun, Zhe, Zhiwei, Zhirui, Ziming. For our "UVABasketball Team Group" members, I am grateful for the times we played basketball together. It was great fun. For our "Eating Group" members, I would like to express my appreciation for the wonderful times we shared every day during our meals. I am grateful for the collective fun of our "Easter Trip Group" and to our "drivers" (Tao, Teng). Thank you, Wangyuan and Yunlu, for bringing me items from our home country. I appreciate Qi Wang's academic and personal assistance. Thanks to Teng, Yixian, and Kefan for teaching me how to swim, even though I haven't quite mastered it yet. I am grateful to Zhiwei, Ziming, and Xiaowei for leaving behind their furniture for me to use. Finally, thank you, Aozhu, for bringing joy to our lives.

Thank you to my closest neighbors (*listed alphabetically by the first capital letter*): Dan, Dongwei, Hui, Jianghua, Jiayun, Jin, Jun, Kefan, Kexin, Ming, Shiqi, Tao, Wenxi, Xiaotian, Xiaowei, Yangjun, Yue & Zhirui, Yunhua. I appreciate the trust, connection, and mutual assistance we have shared in our daily lives. Thanks to our "Let's Paddle Together Group". We often gather together for social events, playing board games and enjoying each other's company. Thank you for the delicious food and great memories. I am grateful for the trust and connection we have in our daily lives. We often gather together and enjoy playing board games and video games. I appreciate Yunhua's unwavering help and sharing, Yue & Zhirui and Jianghua's food, sharing, conversations, and support. Thank you, Jiayun and Kefan, for every gathering, outing, and assistance. I cherish the meal times with Jun. Thanks to Jin for the delicious food and the lovely Ginger. I am grateful

to Shiqi and Xiaoqi. Thank you, Xiaotian, Xuelai, Ming, Kexin, Wenxi and Hui for sharing food with me.

I would like to express my gratitude to my former teachers, Chunliang Wang and Haibo Li, as well as Bin Wang. Additionally, I want to thank my good friends Yuqi, Kaijie, Anlin, Mandy, Haisheng, Facheng, Yisong, Zixiu, Hao, Ning, Fangxiao, Fang, Xin, Yixin, Naifan and my previous roommates, Chao, Yulin, Yu, Zhe, and Wentao.

Of course, I would like to extend a special and heartfelt thank you to Dongwei & Yue, who have provided unwavering and lifelong support without any hesitation.

Lastly, I would like to express my gratitude to my family, especially my parents and my girlfriend, for their unwavering love and support. My girlfriend, Chenlei, has always supported and encouraged me without reservation, even during my darkest times. I am grateful for the care and concern shown by her and her family. I am grateful to my girlfriend for drawing the cover of my dissertation cover with full of love. If life is like a solitary journey across a vast ocean, then you are the guiding light in the darkness. With an endless night ahead, your light accompanies me. I hope we can continue this journey together. I am thankful for my parents, who have given me everything and are my greatest support. The love of parents surpasses all praise and triumphs over everything else.