Visual Kinship Recognition of Families in the Wild

Joseph P. Robinson, *Student Member, IEEE,* Ming Shao, *Member, IEEE,* Yue Wu, Hongfu Liu, *Student Member, IEEE,* Timothy Gillis, *Student Member, IEEE,* and Yun Fu, *Senior Member, IEEE*

Abstract—We present the largest database for visual kinship recognition, *Families In the Wild* (FIW), with over 13, 000 family photos of 1, 000 family trees with 4-to-38 members. It took only a small team to build FIW with efficient labeling tools and work-flow. To extend FIW, we further improved upon this process with a novel semi-automatic labeling scheme that used annotated faces and unlabeled text metadata to discover labels, which were then used, along with existing FIW data, for the proposed clustering algorithm that generated label proposals for all newly added data– both processes are shared and compared in depth, showing great savings in time and human input required. Essentially, the clustering algorithm proposed is semi-supervised and uses labeled data to produce more accurate clusters. We statistically compare FIW to related datasets, which unarguably shows enormous gains in overall size and amount of information encapsulated in the labels. We benchmark two tasks, kinship verification and family classification, at scales incomparably larger than ever before. Pre-trained CNN models fine-tuned on FIW outscores other conventional methods and achieved state-of-the art on the renowned KinWild datasets. We also measure human performance on kinship recognition and compare to a fine-tuned CNN.

Index Terms—Large-Scale Image Dataset, Kinship Verification, Family Classification, Semi-Supervised Clustering, Deep Learning.

1 Introduction

Visual kinship recognition has an abundance of practical uses, such as issues of human trafficking and in missing children, problems from today's refugee crises, and social media platforms. Use cases exist for the academic world as well, whether for machine vision (*e.g.*, reducing the search space in large-scale face retrieval) or a different field entirely (*e.g.*, historical & genealogical lineage studies). However, to the best of our knowledge, no reliable system exists in practice. This is certainly not due to a lack of effort by researchers, as many works focused on kinship [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26].

Challenges preventing visual kinship recognition from transitioning from research to reality are two-fold:

- Existing data resources for visual kinship are too small to capture true data distributions.
- Hidden factors of visual similarities/differences between blood relatives are complex and less discriminant than in other more conventional problems (*e.g.*, object classification or even facial identification).

Clearly, a large image-set that properly represents families worldwide is needed, which also meets the capacity of more complex, data-driven models (*i.e.*, deep learning), thus, motivating us to build the first large-scale image database for kinship recognition, *Families In the Wild* (FIW). FIW is made-up of rich label information that captures the complex, hierarchical structures of 1,000 unique family trees. Families consist of an average of about 13 photos of each (*i.e.*, over 13,000 family photos), and family sizes range from 3-38 members, with most subjects having multiple samples at various ages (see Fig. 1). FIW is the **largest** and **most comprehensive** database of its kind.[1]

Deep learning can now be applied to the problem, as we demonstrate on two benchmarks, kinship verification and family classification. We fine-tune deep models to improve all benchmarks, and provide details on the training procedure. We also measure human performance on verification and compare with benchmarks.

We use a multi-modal labeling model to optimize the annotation process. This includes a novel semi-supervised clustering method that works effectively in practice (*i.e.*, generates label proposals for new data using existing labeled data as side information). For this, we increase the amount of available side information using existing labels (*i.e.*, names), labeled faces, and text metadata collected with the family photos. We show a significant reduction in manual labor and time spent on labeling new data.

Families In the Wild (FIW) was first introduced in [27]. This work adds to the previous work in a number of ways. Listed here are contributions made in this journal extension:

1) Added additional faces for verification and complete families for classification (Section 3).
2) Improved the labeling process with novel semi-supervised clustering method (Section 4).
3) Boosted baseline scores using up-to-date deep learning approaches (Section 5).
4) Obtained state-of-the-art on smaller datasets via transferring CNN fine-tuned on FIW (Section 5.5).
5) Conducted kinship verification experiment on humans and compared with algorithms (Section 5.6).

- *Joseph P. Robinson, Yue Wu, Hongfu Liu and Timothy Gillis are with the Department of Electrical and Computer Engineering, Northeastern University, MA, 02115, USA, (e-mail: robinson.jo@husky.neu.edu).*
- *M. Shao is with the Department of Computer and Information Science, University of Massachusetts Dartmouth, MA, 02747, USA, (e-mail: mshao@umassd.edu).*
- *Y. Fu is with the Department of Electrical and Computer Engineering and the College of the Computer and Information Science, Northeastern University, MA, 02115, USA, (e-mail: yunfu@ece.neu.edu).*

[1] Download FIW via project page, smile-fiw.weebly.com/.

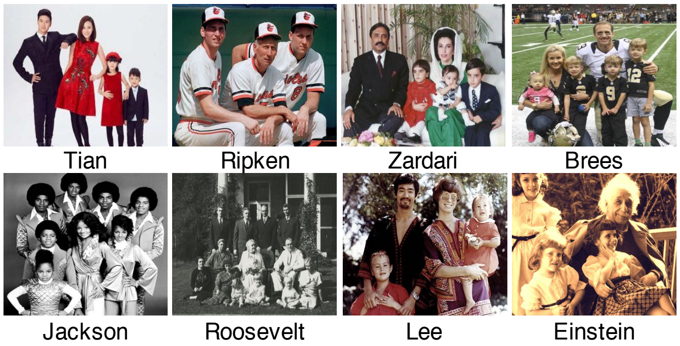


Fig. 1. Photos of families sampled randomly from FIW (*i.e.*, 8 of $1,000$).

2 Related Works

2.1 Related Databases

The story of visual kinship recognition begins in 2010, at which time the first kin-based image collection (*i.e.*, CornellKin) was made public [1]. CornellKin included 150 *parent-child* face pairs (*i.e.*, celebrities and their parents). Next, UB KinFace-I & II [16], [28], [29] was introduced to address a slightly different view of kinship recognition– both young and old faces of parents are paired with a child, with a total of 600 face photos of 400 unique subjects (*i.e.*, celebrities and politicians). Then, KinWild I-II [30] was released and used in a 2015 FG Challenge [31], which too focused on *parent-child* pairs. Shortly thereafter, Family101 [7] was introduced as the first attempt of multi-class classification (*i.e.*, *one-to-many*) for kinship recognition. Thus, it is an organized set of structured families [7], including 206 sets of parents and their children (*i.e.*, *core families*) that make up 101 unique family trees. In 2015, TSKinFace [12] was built to support yet another view of kinship recognition, tri-subject verification, where both parents and a child are used– 513 Father/Mother-Daughter pairs and 502 Father/Mother-Son pairs (*i.e.*, *two-to-one* verification).

However, even after all these contributions, there existed no single resource that satisfied the concerns of insufficient data. A single resource with the features of previous works, but in a more complete and abundant manner, was the underlying vision for FIW. As shown in Tables 1 & 2, and discussed in later sections, FIW far exceeds others in terms of number of families, face pairs, and relationship types.

2.2 Automatic Kinship Recognition

As mentioned, Fang et al. [1] first attempted kinship verification on *parent-child* face pairs. They proposed selecting the 14 (of 44) most effective hand-crafted features. Following this, researchers recognized that a child's face more closely resembles their parents at younger ages [16], [28], [29]. In response, they used transfer subspace learning methods that uses the younger faces of parents to help fill the appearance gap between their older faces and that of their children. To benchmark the KinWild dataset, Lu et al. [32] proposed a metric learning method used in Euclidean space called NRML and its multi-view counterpart (MNRML) that learns a common distance metric for multiple feature types. Fang et al. [7] focused on *one-to-many* (*i.e.*, family classification) by representing faces as a linear combination of sparse features (*i.e.*, feature selection via lasso) of 12 facial parts encoded via a learned dictionary.

Progress made in kinship recognition, along with release of varying task protocols, coincides with an increasing availability of structured and labeled data. Although there have been several significant contributions, none have overcome the challenges posed earlier.

2.3 Deep Kinship Recongition

Since the AlexNet CNN [33] won the 2012 ImageNet Challenge [34], deep learning has achieved state-of-the-art in a wide range of machine learning tasks. Central to this frenzy has been facial recognition [35], [36], [37]. In spite of this, there are only a few works that use deep learning for kinship recognition [38], [39], [40], [41].

Deep learning has yet to show an advantage for visual kinship recognition, with metric learning seeming more promising. As mentioned in a recent literature review [42], the reason for this is due to insufficient amounts of data. In this work, we include several benchmarks on FIW using deep learning, obtaining a clear advantage in both tasks.

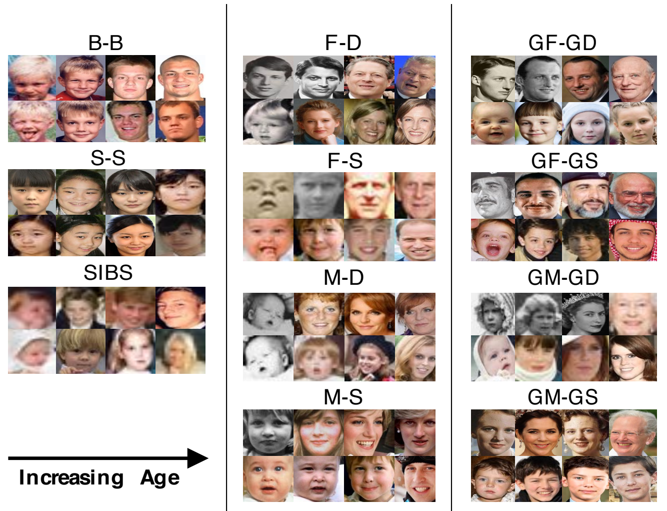
2.4 Semi-Automatic Image Tagging & Data Exploration

Automatic image tagging was recently done by first labeling a small amount of the data, and then using it as side information to help guide the clustering process in a semi-supervised manner [43]. Following this, we take advantage of side information from labeled FIW.

Previous works used image captions, whether from Flickr or other sources of images tagged by users, to discover labels and annotate images in an automatic fashion [44]. Generally, methods mining text for image tags treat it as a problem of noisy labels [45], [46]. CASIA-WebFace [47], a large-scale dataset for facial recognition, successfully extended the scale of the renowned LFW [48]. By crawling the web, and leveraging knowledge from IMDB, multiple face samples for $10,000$ unique subjects were collected. Although related in the sense of automatic labeling, these problems are very different from the one we present here. We aim to add more data to underrepresented families of the FIW database, and doing so by using the existing labels for each family as side information to guide our semi-supervised clustering method. We wish to maximize the number of labeled faces available to facilitate the clustering in order to generate label proposals. For this, we use the existing FIW labeled faces and the text metadata of the unlabeled data to automatically tag faces using an iterative process governed by both visual and contextual evidence. As discussed in Section 5.4, our method consistently improves with increasing amounts of side information.

3 Families in the Wild (FIW) Database

We next cover the FIW database. First, we review the existing FIW and old labeling scheme [27]. Then, we introduce the improved semi-automatic labeling process. Finally, we compare the two.



(a) Same Generation (b) 1^{st} Generation (c) 2^{nd} Generation

Fig. 2. Samples of 11 pair types of FIW. Each type is of a unique pair randomly selected from a set of diverse families to show variation in ethnicity, while four faces of each individual depict age variations.

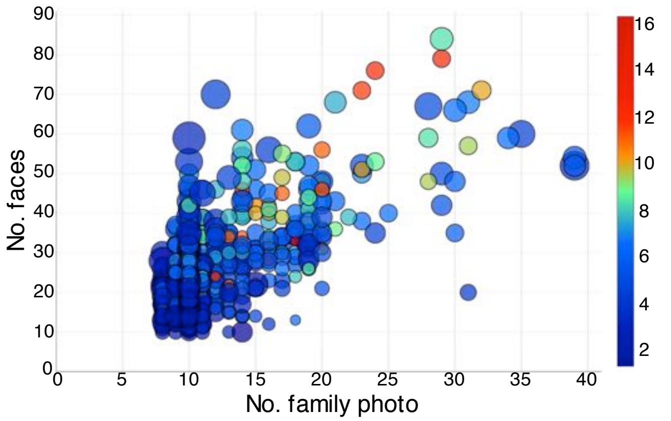


Fig. 3. Database statistics: Horizontal and vertical axes represent counts for photos and faces per family, respectively. Bubble size and color represent counts for members and average faces per member, respectively.

3.1 Existing FIW

Our goal for FIW was to collect about 10 family photos for 1,000 unique families and support with 2 types ground-truth labels, photo-level (*i.e.*, who is where in the image) and family-level (*i.e.*, all members and the relationships between them). Fig. 4 depicts the 2 label types. FIW is organized as follows: each family is assigned a unique ID (*i.e.*, FID), and pictures collected are also assigned a unique ID (*i.e.*, PID). Finally, members added are assigned their own unique ID (*i.e.*, MID). For instance, $FID_1 \rightarrow MID_1$ in PID_1 refers to the first member of the first family in the first photo collected. The order of IDs is arbitrary, as assignments were made in the order that the family, member, and photo were added. Before introducing the new and improved semi-automatic process, we briefly review the process used initially in [27], which involved 3 steps: (1) *Data Collection*, (2) *Data Labeling*, and (3) *Data Parsing*.

For *Data Collection*, a team of 8 students from different parts of the world and with vast knowledge of famous persons, compiled a list of families and collected images with a primary focus on their place of origin (*i.e.*, an attempt to compile a diverse family list). Table 3 lists the ethnicity distributions of the 1,000 families. Note that this is not the exact distribution, as each family is counted once according to the *root* member for which the search was based (*i.e.*, not per member, but per family). For instance, for Spielberg's family we consider just Stephen. Future work could entail adding more families from underrepresented ethnic groups, as the distribution still favors Caucasians.

For *Data Preparation*, we built a labeling tool to guide the process of generating the two label types. Labelers would work through all family photos on a family-by-family basis, specifying who is in each photo by clicking member faces and choosing their names from a drop-down menu. Names, genders, and relationships for members were only entered on the first image they are present–once added to the family the labelers just selected their names each time they appeared in a photo.

For *Data Parsing*, all family photos were detected using classic HOG features trained on top of a linear classifier using image pyramids and sliding windows via DLIB [49]. Faces were cropped and normalized as done in [50], and then resized to 224 × 224. Finally, the structure of the database was organized into a hierarchy of directories, FID→MID→Face-ID (*i.e.*, 1,000 folders, $F0001$-$F1000$, containing family labels and folders for MIDs with face samples of that member).

Even though it only took a small team to label 10,676 family photos and 1,000 families, the process relied heavily on human input. Plus, in the end, many families were not properly represented (*i.e.*, either too few members, face samples, or family photos). Thus, we aim to reduce the manual labor and overall time requirements to add additional data provided various amounts of labels existed for each (*i.e.*, 61 existing families and 4 replacement). We added replacement families (*i.e.*, newly added families) to make up for cases of overlapping families or an insufficient online presence when searching for photos (*i.e.*, unable to locate family photos for 2 of the under-represented families). Before we propose the semi-automatic labeling model, we first review the two benchmarks included in this work, along with the related statistics of each. We then present the new labeling process that enabled us to add additional data with far less manual labor and in just a fraction of the time.

3.2 Data Preparation

Due to the nature of the label structure, FIW can serve as a resource for various types of vision tasks. For this, we benchmark the two popular tracks, kinship verification and family classification. Next, we introduce both of these tasks and the means of preparing the data.

3.2.1 Kinship verification

Kinship verification aims to determine whether two faces are blood relatives (*i.e.*, kin or non-kin). Prior research mainly focused on *parent-child* pairs (*i.e.*, father-daughter (F-D), father-son (F-S), mother-daughter (M-D), and mother-son (M-S)); some considered sibling pairs (*i.e.*, brother-brother (B-B), sister-sister (S-S), and brother-sister/mixed

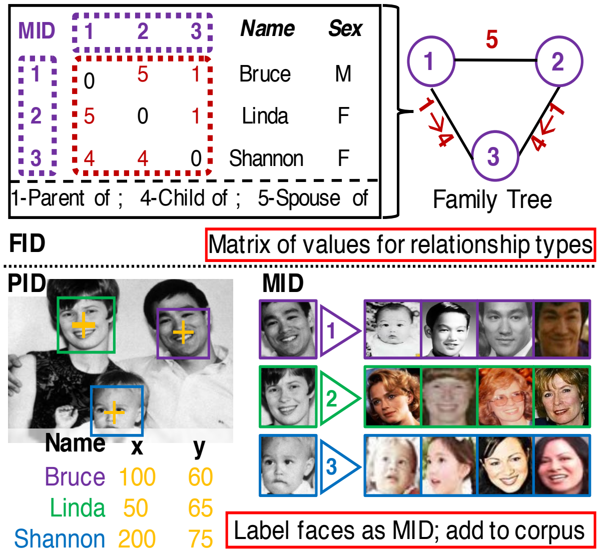


Fig. 4. Visual of the 2 label types of FIW, *Family-level* (FID) and *Photo-level* (PID). FID has individual family member (MID) and relationship information. PIDs contain information of MIDs + their locations in photos.

gender siblings (SIBS)). However, research in both psychology and computer vision revealed that different kin relations render different familial features, which motivated researchers to model different relationship types independently. With the existing image datasets used for kinship verification limited to, at most, 1,000 faces and typically only 4 relationship types, we believe such minimal data leads to overfitting and, hence, models that do not generalize well to unseen data captured *in the wild*. FIW currently supports 11 relationship types (see Fig. 10), 4 being introduced to the research community for the first-time (*i.e.*, *grandparent-grandchild*) and, most importantly, each category is made up of many more pairs– 418,000 face pairs in [27] has increased to 656,954 after extending FIW via the proposed semi-supervised approach.

The 11 relationship types provide a more accurate representation for real-world scenarios. As mentioned, FIW is structured such that the labels can be parsed for different types of tasks and experiments, and additional kinship types can easily be inferred.

3.2.2 *Family classification*

Family classification aims to determine the family an unknown subject belongs to. Families are modeled using the faces of all but one family member, with the member left out used for testing. This *one-to-many* classification problem is a challenging problem that gets more challenging with more families. This is becuase families contain large intra-class variations that typically fool the feature extractors and classifiers, and each additional family further adds to the complexity of the problem. Additionally, and like conventional facial recognition, when the target is unconstrained faces *in the wild* [48] (*e.g.*, the variation in pose, illumination, expression, etc.), the problem continues to become more difficult. In [27], the experiment included only 316 families (*i.e.*, families with 5+ members). In this extended version, we now can include 524 families with the added data. We next present the process followed to extend FIW.

3.3 Extending FIW

We set out to use the proposed semi-supervised model to generate label proposals for new data while using existing labels as side information to yield more accurate clusters. As explained in Section 5.4 and shown in Fig. 7, the proposed clustering method only improves with more side information. Thus, we want to maximize the amount of side information (*i.e.*, labeled faces) available. We do this by inferring highly confident labels by aligning faces and names from the unlabeled photos and corresponding text metadata. In addition, and when available, we model labeled data to discriminate between family members in a photo. In the end, clusters are saved as ground truth upon being verified by a human. Significant savings, in labor and time costs, resulted from using this labeling process.

A single family is processed at a time to reduce both the search and label spaces. We aim to discover labels with evidence from multiple modalities (*i.e.*, visual and contextual information). This not only increases the amount of side information available for clustering, but also the sample count to use for discovering more labels as face-name pairs. The cluster assignments (*i.e.*, label proposals) were then manually inspected (*i.e.*, validated).

We demonstrate the effectiveness of the new labeling scheme by comparing the number of user inputs (*i.e.*, mouse clicks and keystrokes) and overall time with the process followed in [27]. It took just a few inputs and a few minutes on average per family, opposed to hundreds of inputs and several minutes to over an hour (see Table 2).

We next explain the improved multi-modal scheme made-up of 4 steps: (1) *Data Collection*, (2) *Data Preparation*, (3) *Label Generation*, and (4) *Label Validation*. The goal of (1) and (2) is to gather and increase the amount of side information available for (3), while (4) is to ensure correct labels for all new data. In other words, we set out to increase the labeled sample pool (*i.e.*, side information) by inferring labels for unlabeled faces, which adds to the set of training exemplars. The faces that are still unlabeled in (3) are clustered using all labels as side information. All newly added data is then verified by a human. The process is illustrated in Fig. 5, which we next describe step-by-step.

3.3.1 *Step 1: Data Collection*

The goal is to collect additional data for under-represented families of FIW. These families lacked in the number of members, faces, and/or family photos. In total, there were 65 families that were extended, with 1 family replaced due to a lack of available data, and 3 other overlapping families that were merged (*i.e.*, Catherine, Duchess of Cambridge, along with her immediate family, merged with the *Royal* family, as her spouse Prince William share 2 children and, thus, the two families (*i.e.*, sets of in-laws) are connected by kinship). Several new labels and relationships resulted from this merge, as the *Royal* family went from having 29 to 38 members, which is the largest family of FIW. Database statistics are described in Fig. 3 and Table 3.

There were 2 requirements for the data collection: (1) rich text metadata that described the subjects in the photos and (2) at least 1 portrait face (or profile picture) for new family members. These requirements are for *Step 2*, as the label

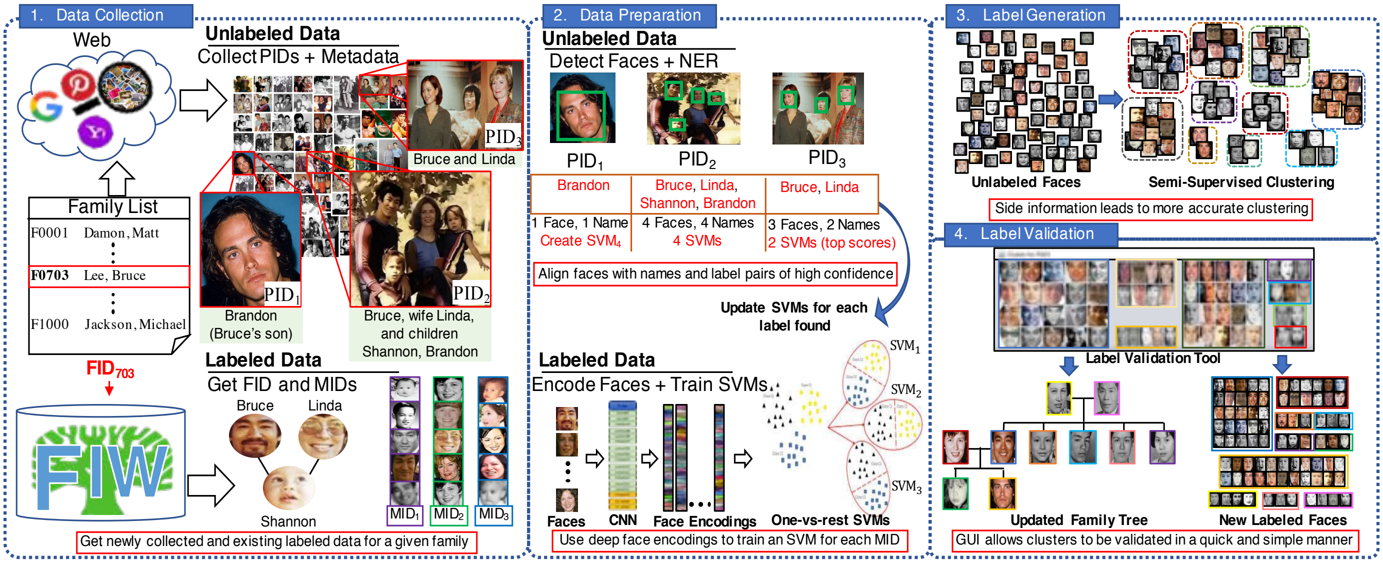


Fig. 5. Semi-automatic labeling pipeline. *Data Collection.* Photos and text metadata are collected for underrepresented families in FIW and assigned unique IDs (*i.e.*, PIDs). Each new member requires at least 1 profile picture (*e.g.*, Brandon in PID_1) to add to known labels. *Data Preparation.* With the existing FIW labels, we next aim to increase the amount, both in labeled faces and member labels, using multiple modalities– names in metadata and scores of SVMs are used to automatically label some unlabeled data– face-name pairs were assumed labeled for cases of high confidence. Starting from profile pictures (*i.e.*, 1 face, 1 name) and working towards less trivial scenarios (*e.g.*, 3 faces and 2 names, with 2 faces from 1 member at different ages, like in PID_3). This step adds to the amount of side information used for clustering. *Label Generation.* Label proposals for remaining unlabeled faces are generated using the proposed semi-supervised clustering model that leverages labeled data as side information to better guide the process. *Label Validation.* A GUI designed to validate clusters and ensure clusters are matched to the proper labels.

TABLE 1
Face pair counts for FIW and other kinship image collections, showing FIW far outdoes all others in the 7 *sibling* & *parent-child* pair types. Plus, introduces 4 *grandparent-grandchild* types for the 1^{st} time. Table 2 further characterizes FIW and Fig. 10 shows samples for each pair type.

	siblings			parent-child				grandparent-grandchild				Total
	B-B	S-S	SIBS	F-D	F-S	M-D	M-S	GF-GD	GF-GS	GM-GD	GM-GS	
KinWild I [32]	0	0	0	134	156	127	116	0	0	0	0	533
KinWild II [32]	0	0	0	250	250	250	250	0	0	0	0	1,000
Sibling Face [51]	232	211	277	0	0	0	0	0	0	0	0	720
Group Face [8]	40	32	53	69	69	62	70	0	0	0	0	395
FIW(Ours) [27]	**103,724**	**39,978**	**73,506**	**92,088**	**129,846**	**82,160**	**112,618**	**7,078**	**4,830**	**6,512**	**4,614**	**656,954**

expansion is done for all new members by mining the trivial face-name pairs of portrait photos– portraits can align the single face to the single name with high confidence (more details provided in the following step).

3.3.2 Step 2: Data Preparation

We now aim to maximize the amount of side information available for the clustering process. For this, we took advantage of both labeled (*i.e.*, faces and names) and unlabeled data (*i.e.*, detected faces and text metadata) in order to automatically infer labels for some unlabeled faces (see *Data Preparation* in Fig. 5). We next break down the different components used in this step for clarity.

Text metadata (*i.e.*, image captions) are collected along with all photos in *Step 1*. For each family, all metadata gets processed using a Name Entity Recognition (NER) classifier [52] to output a list of detected names. Then, a *Look-Up-Table* (LUT) of possible names for each member is generated– *i.e.*, keys are member IDs (MIDs) and values are possible references to that member (*e.g.*, *Bruce* aka *Bruce Lee* aka *Brandon's father*). One challenge is from name variations (*e.g.*, a person legally named *Joseph* might be called *Joe*); additionally, there are name titles (*e.g.*, *Queen Elizabeth II* might be called *Elizabeth II*, *Elizabeth*, or, in older photos, *Princess Elizabeth*). Nicknames, which are less trivial, pose additional challenges (*e.g.*, *Robert Gronkowski* called *Rob*, *Robert*, or by his nickname *Gronk*). To address this, we first generated a LUT with detected references to each subject, and then augmented the search using additional tags (*e.g.*, adding titles and last names). The LUTs are later used to find evidence in the text metadata of family members' presence in a photo.

New MIDs found in profile photos (*e.g.*, PID_1 in Fig. 5)– when processing a family, each image that has a single face detected and just one name in its metadata is considered a profile photo. Profile photos are first to be processed. The name detected in the metadata is compared to all names for members stored in the LUTs. If there are no matches, the subject is then added as a new member in that family. A LUT of names is then generated for each new member, and the name of highest frequency (*i.e.*, number of detections in

TABLE 2
Comparison of FIW with related datasets.

Dataset	No. Family	No. People	No. Faces	Age Varies	Family Trees
CornellKin [1]	150	300	300	×	×
UBKinFace [16], [29]	200	400	600	✓	×
KFW-I [30]	×	533	1,066	×	×
KFW-II [30]	×	1,000	2,000	×	×
TSKinFace [12]	787	2,589	×	✓	✓
Family101 [7]	101	607	14,816	✓	✓
FIW [27]	**1,000**	**10,676**	**30,725**	✓	✓

all metadata) recorded as the name corresponding to their assigned MID (*e.g.*, MID_6 for the sixth member).

Unlabeled and labeled faces are encoded as $4,096D$ features from the fc_7-layer of the pre-trained VGG-Face CNN model [36]. *One-vs-rest Support Vector Machine* (SVM) models are trained for each member using labeled samples from all other members of that family as the negatives. Next, profile photos are processed (*i.e.*, 1 name and 1 face). Names that match an existing label are added to corresponding MID data pools, while mismatched names are added as a new MID with a LUT generated. This shows the benefit of including profile pictures for each new member, which makes it so all family members are known. It is important to note that SVMs are updated each time a new labeled face is added.

Discovering labels continues in a similar fashion, except now the SVMs play a more critical role. Now moving on to images with 2 faces and 2 names, the 2 SVMs of the respective members are used to classify the 2 faces. Provided high scores and no conflicts, labels were inferred. Cases with low confidence or conflicts were skipped, leaving those faces to be labeled via clustering. Next, photos with 3 faces and 3 names are processed, then 4 faces and 4 names, and so on and so fourth. After all one-to-one cases are processed, photos with a different number of names and faces are processed. For each photo, only SVMs that correspond to a LUT with matching names are used. Thus, justifying a requirement of *Step 1*– collect rich metadata in terms of specifying members present in photos.

It should also be noted that some families benefited far more than others in this process. Nonetheless, roughly 25% of the $2,973$ added faces were correctly labeled by this simple multi-modal process.

TABLE 3
Ethnicity distribution of FIW. *Mix* are families with $2+$ ethnicities (*e.g.*, Bruce Lee (*Asian*) and wife Linda (*Caucasian*) with 2 children (*Mix*).

Caucasian	Spanish/Latino	Asian	African/AA	Arabic	Mix
64%	10.7%	9.1%	8.2%	2.0%	6.0%

3.3.3 Step 3: Label Generation

Label proposals were generated for unlabeled faces using the proposed semi-supervised clustering method. To get the most out of our model we automatically labeled additional data in *Step 2*, while identifying all new members being added to each family. Hence, the number of members (*i.e.*, k) is known for each family.

More details, including the objective function and solution, are provided in Section 4.

3.3.4 Step 4: Label Validation

Finally, clusters (*i.e.*, labels) are validated by a human. This is a three-part process: assign an MID to each cluster; validate each cluster, which is displayed in a grid of faces in the order of confidence score; specify gender and relationships of newly added members. As shown in Fig. 5, a JAVA interface was designed to generate ground-truth for new data with just a few clicks of the mouse and minimal time per family. The inputs are cluster assignments for a family, with faces listed in order of confidence (*i.e.*, cosine distance from centroid). MIDs were assigned in *Step 2* (*i.e.*, inferred from text, SVM scores, or both), which must also be validated. The outputs are labels for each PID and an updated relationship matrix (Fig. 4).

3.3.5 Discussion

Seven families of various sizes were used to compare the old [27] and proposed labeling schemes– old scheme took 4,124 inputs in about 2.75 hours, and just 95 inputs in about 18.1 minutes via the new (see Table 4). Collecting and labeling the data for the extended FIW was done by a single person in days; it initially took a small team several months with the old scheme. Thus, demonstrating a significant savings in manual labor and time (note that greater amounts of data was originally collected, however, relative savings in time and manual labor clearly yields from process used in this extended version). A possible future direction is to use this scheme to extend families of FIW with video data. Another possibility is to use this method to extend the number of families, which, if on the order of thousands or more, then automating *Step 1* could further reduce savings (*i.e.*, web scrape for family information (*e.g.*, *Wiki*) and photos (*e.g.*, *Google*, *Bing*, etc.)).

4 Semi-Supervised Face Clustering

Labeling is a human-necessary and expensive task to benchmark data sets. Here we aim to accelerate the process by using some labeled data in advance. In this part, we demonstrate a novel semi-supervised clustering for labeling. Let $X = \{x_i\}$ be the data matrix with n instances and m features and S be a $n' \times K'$ side information matrix, which denotes n' labeled data instances into K' classes. Our goal is to make use of S to guide the remaining instances into K classes, where $K' \leq K$.

4.1 Objective Function

Inspired by our previous work [43], [53], a partition level constraint is used to make the learnt partition agree with partial human labels as much as possible. To demonstrate

TABLE 4
Previous (white) vs new (shaded) labeling processes compared in terms of inputs (keyboard and mouse clicks) and time (hours:minutes:seconds).

	Bruce Lee	Michael Jordan	John Malone	Craig Mccaw	Marco Reus	British Royal	Michael Jackson	Total
Inputs (count)	551	97	153	178	35	1,838	1,272	4,124
Inputs (count)	**12**	**6**	**10**	**15**	**7**	**21**	**24**	**95**
Time (h:m:s)	0:15:08	0:5:31	0:5:18	0:6:16	0:4:24	1:25:23	0:44:52	2:46:52
Time (h:m:s)	**0:1:11**	**0:0:31**	**0:1:05**	**0:0:56**	**0:0:31**	**0:6:44**	**0:7:13**	**0:18:11**

the effectiveness of our labeling mode, K-means with cosine similarity is employed as the core clustering method to handle high-dimensional data due to its high efficiency and robustness. The following is our objective function,

$$\min \sum_{k=1}^{K} \sum_{x_i \in \mathcal{C}_k} f_{cos}(x_i, m_k) + \lambda U_c(S, H \otimes S), \tag{1}$$

where f_{cos} is the cosine similarity, H is the final partition, $H_S = H \otimes S$ is part of H which the instances are also in the side information S, m_k is the centroid of $\mathcal{C}_k$, U_c is the well-known Categorical Utility Function [54] and λ is the trade-off parameter.

To better understand the last term in Eq. 1, we give the detailed calculation of U_c. Given two partitions S and H_S containing K' and K clusters, respectively. Let $n_{kj}^{(S)}$ denote the number of data objects belonging to both cluster $C_j^{(S)}$ in S and cluster C_k in H_S, $n_{k+} = \sum_{j=1}^{K'} n_{kj}^{(S)}$, and $n_{+j}^{(S)} = \sum_{k=1}^{K} n_{kj}^{(S)}$, $1 \leq j \leq K'$, $1 \leq k \leq K$. Let $p_{kj}^{(S)} = n_{kj}^{(S)}/n'$, $p_{k+} = n_{k+}/n'$, and $p_{+j}^{(S)} = n_{+j}^{(S)}/n'$. We then have a normalized contingency matrix (NCM), based on which a wide range of utility functions can be accordingly defined. For instance, the widely used category utility function can be computed as follows:

$$U_c(H_S, S) = \sum_{k=1}^{K} p_{k+} \sum_{j=1}^{K'} (\frac{p_{kj}^{(S)}}{p_{k+}})^2 - \sum_{j=1}^{K'} (p_{+j}^{(S)})^2. \tag{2}$$

It is worthy to note that U_c measures the similarity of two partitions, rather than two instances. The larger value of U_c indicates the higher similarity.

4.2 Solution

We notice that the first term in Eq. 1 is the standard K-means with cosine similarity. Could we still apply K-means optimization to solve the problem in Eq. 1? The answer is yes! Due to our previous work [55], we provide a new insight of U_c by the following lemma.

Lemma 1. *Given a fixed partition S, we have*

$$U_c(H_S, S) = -||S - H_S G||_{\mathrm{F}}^2 + \text{constant}, \tag{3}$$

where G is the centroid matrix of S according to H_S.

By the above lemma, the second term in Eq. 1 can also be transformed into a K-means problem with squared Euclidean distance. Then a K-means-like algorithm can be used on the augmented matrix with modified distance function and centroid update rule for the final partition.

First an augmented matrix D is introduced as follows.

$$D = \begin{bmatrix} X_S & S \\ X_T & 0 \end{bmatrix} \text{ with } X = \begin{bmatrix} X_S \\ X_T \end{bmatrix}, \tag{4}$$

where d_i is the i^{th} row of D, which has of two parts, $d_i^{(1)}$ and $d_i^{(2)}$ (*i.e.*, $d_i^{(1)} = (d_{i,1}, \cdots, d_{i,d_m})$ presents the feature space and $d_i^{(2)} = (d_{i,d_m+1}, \cdots, d_{i,d_m+K'})$ denotes the label space). Zeros in D are the artificial elements, rather than the true values so that all zeros contribute to the computation of the distance and centroids, which inevitably interfere the cluster structure. To make the zeros in D not involved in the calculation, we give the new update rule for the centroids of D. Let $m_k = (m_k^{(1)}, m_k^{(2)})$ be the k^{th} centroid $\mathcal{C}_k$ of D, we modify the computation of centroids as follows.

$$m_k^{(1)} = \frac{\sum_{d_i \in \mathcal{C}_k} d_i^{(1)}}{|\mathcal{C}_k|}, \;\; m_k^{(2)} = \frac{\sum_{d_i \in \mathcal{C}_k} d_i^{(2)}}{|\mathcal{C}_k \cap X_S|}. \tag{5}$$

and the distance function is also adjusted as

$$f(d_i, m_k) = f_{cos}(d_i^{(1)}, m_k^{(1)}) + \mathbf{1}(d_i \in S) f_{sqE}(d_i^{(2)}, m_k^{(2)}), \tag{6}$$

where $\mathbf{1}$ returns 1 if the condition is satisfied, otherwise 0.

The correctness and convergence of the modified K-means is similar to one in [43].

5 Experiments

We conduct the following experiments: benchmark kinship verification and family classification; evaluate the proposed semi-supervised clustering method at the core of the new labeling scheme; fine-tune CNNs using FIW and evaluate on KinWild I & II (*i.e.*, transfer-learning); measure human performance on kinship verification and compare to top scoring algorithms.

The subsequent subsections are organized as follows. First, we review the visual features, metric learning methods, and deep learning that is common in all experiments. Then, we dive into the experiments mentioned above. We introduce each independently, but with the same structure: experimental settings, experiment-specific training philosophy, and then the results and analysis.

5.1 Experimental Setting

For the sake of organization, all low-level features and metric learning approaches used throughout are listed and described in this section. Most are in two or more experiments, however, even those used for verification, for example, are still treated as common information, and thus is described

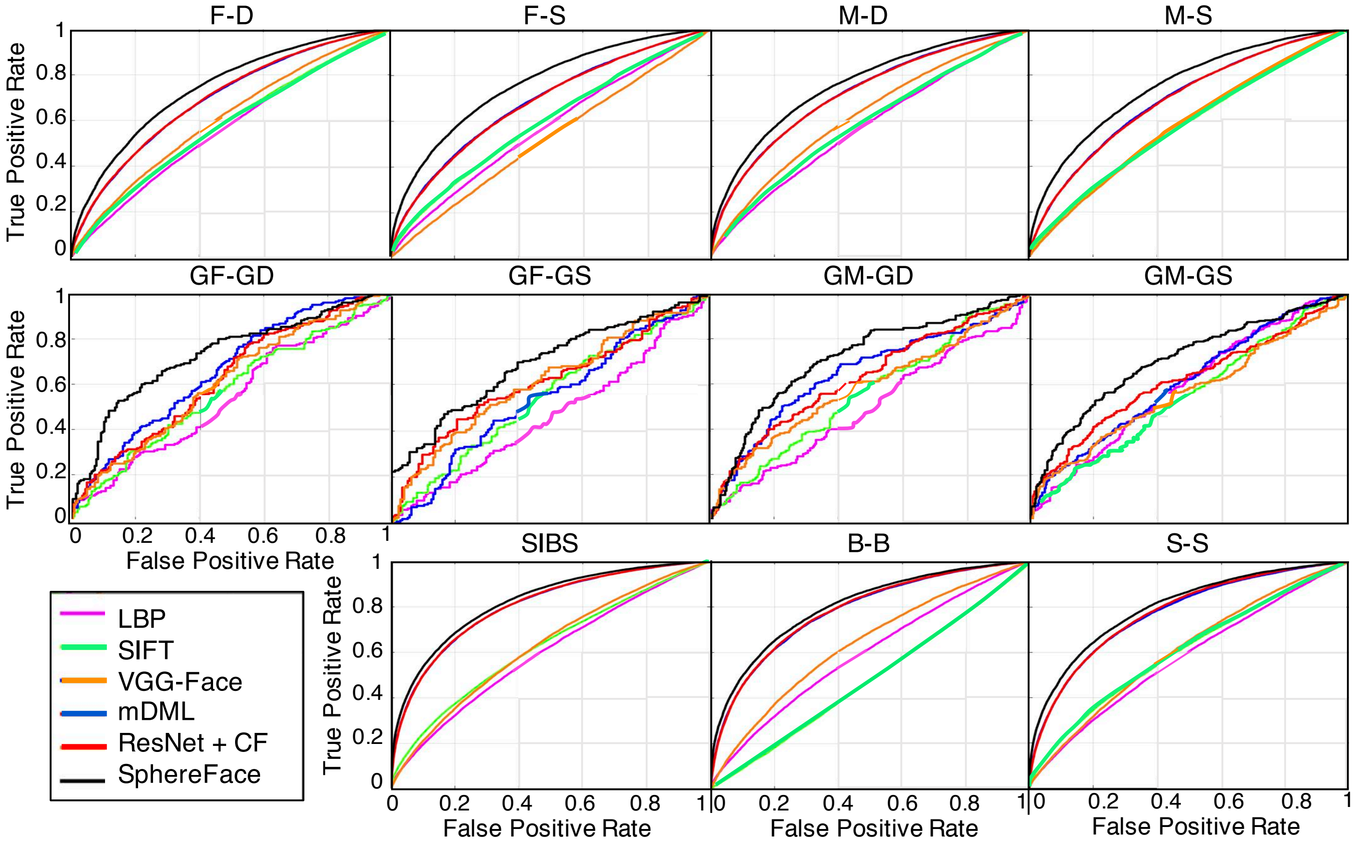


Fig. 6. Relationship type specific ROC curves.

alongside other items of preliminary information. Following concepts pertaining to "shallow" vision methodology, we review specifications of the pre-trained CNNs used as off-the-shelf feature extractors.

5.1.1 Feature Representations

Detected and aligned faces were normalized and encoded using low-level and CNN-based features. We next describe the descriptors used in this work– SIFT, LBP, pre-trained VGG-Face and ResNet CNNs– each having been widely used in visual kinship and facial recognition problems.

SIFT [66] is amongst the most widely used feature type in object and face recognition. Here we follow the settings of [30]: resize images to 64×64, then extract features from 16×16 blocks with a stride of 8 (*i.e.*, 49 blocks that yields $128 \times 49 = 6,272D$ face feature).

LBP [56] are renown for its effectiveness in tasks such as texture analysis and face recognition. We again follow the settings of [30]: resize images to 64×64, divide into 16×16 non-overlapping blocks, and use a radius of 2 and sampling number of 8. Each block is represented as a $256D$ histograms (*i.e.*, $256 \times 16 = 4,096D$ face encoding).

VGG-Face [36], a pre-trained CNN with the topology of VGG-16: made-up of small convolutional kernels (*i.e.*, 3×3) with a convolutional stride of 1 pixel. VGG-Face is trained on $2.6M$ face images of $2,622$ different celebrities. VGG has worked well on various face databases– 97.3% in accuracy on *YouTube Faces* [67]; 98.95% accuracy on *Labeled Faces in the Wild* [68]. By removing the top two layers– softmax and last fully-connected layer (aka fc8-layer or fc_8)– the CNN can be used as an *off-the-shelf* face encoder [69]. Thus, models get trained on an auxiliary resource and employed on target data. Here, we fed faces through to the fc7-layer (aka fc_7), yielding a $4,096D$ face encoding.

ResNet-22 [58] is a 22-layer residual CNN trained on CASIA-Webface [47]. ResNet-22 has a different network topology than VGG (*i.e.*, more layers made possible via skipping connections in residual blocks to ensure that the signal stays intact by superimposing an identity tensor). Faces were fed through to layer fc_5 ($512D$ encoding).

5.1.2 Metric Learning

Metric learning is commonly used and, sometimes, designed for kinship problems. Four metric learning and graph embedding methods used previously for face-based problems are include: Information theoretic metric learning (ITML) [59], Discriminative Low-rank Metric Learning (DLML) [63], Locality Preserving Projections (LPP) [60], and Large Margin Nearest Neighbor (LMNN) [61].

5.1.3 Deep Learning

Fine-Tuned CNNs. Centerface (CF) [58] loss enhances the discriminative power of deeply learned features by adding a supervision signal to reduce the intra-class variations. SphereFace uses an angular softmax loss, and has most recently claimed state-of-the-art in facial recognition [65]. We fine-tune both these CNNs on FIW.

TABLE 5
Averaged verification accuracy scores (%) for 5-fold experiment on FIW. Note that there was no family overlap between folds.

	siblings			parent-child				grandparent-grandchild				
Method	B-B	S-S	SIBS	F-D	F-S	M-D	M-S	GF-GD	GF-GS	GM-GD	GM-GS	Acc. ± Std.
LBP [56]	55.52	57.49	55.39	55.05	53.77	55.69	54.65	55.79	55.92	54.00	55.36	55.33 ± 1.01
SIFT [57]	57.86	59.34	56.91	56.37	56.24	55.05	56.45	57.25	55.35	57.29	56.74	56.80 ± 1.17
ResNet-22 [58]	65.57	69.65	60.12	59.45	60.27	61.45	59.37	55.37	58.15	59.74	59.70	61.34 ± 3.81
VGG-Face [36]	69.67	75.35	66.52	64.25	63.85	66.43	62.80	62.06	63.79	57.40	61.64	64.89 ± 4.68
+ITML [59]	57.15	61.61	56.98	58.07	54.73	57.26	59.09	62.52	59.60	62.08	59.92	59.00 ± 2.44
+LPP [60]	67.61	66.22	71.01	62.54	61.39	65.04	63.54	63.50	59.96	60.00	63.53	64.03 ± 3.32
+LMNN [61]	67.11	68.33	66.88	65.66	67.08	68.07	66.16	61.90	60.44	63.68	60.15	65.04 ± 3.00
+GmDAE [62]	68.05	68.55	67.33	66.53	68.30	68.15	66.71	62.10	63.93	63.84	63.10	66.05 ± 2.36
+DLML [63]	68.03	68.87	67.97	65.96	68.00	68.51	67.21	62.90	63.96	63.11	63.55	66.19 ± 2.36
+mDML [41]	69.10	70.15	68.11	67.90	66.24	70.39	67.40	65.20	**66.78**	63.11	63.45	67.07 ± 2.44
ResNet+CF [64]	69.88	69. 54	69.54	68.15	67.73	71.09	68.63	**66.37**	66.45	**64.81**	64.39	67.87 ± 2.15
SphereFace [65]	**71.94**	**77.30**	**70.23**	**69.25**	**68.50**	**71.81**	**69.49**	66.07	66.36	64.58	**65.40**	**69.18** ± 3.68

Additionally, we include two state-of-the-art methods based on autoencoders (AE), graph regularized marginalized Stacked AE (GmDAE) [62], and marginalized denoising AE based metric learning (mDML) [41].

5.2 Kinship Verification

Kinship verification is a binary classification problem (*i.e.*, *true* or *false*, aka *kin* or *non-kin*, respectfully). It is the *one-to-one* view of kinship recognition, which is explained next.

5.2.1 Experimental Setting

The protocol we followed is conventional in face-based tasks: 5-fold cross validation with no family-overlap between folds. There are 11 relationship types evaluated (statistics and types shown in Table 1).

For each pair type, we added negative (*i.e.*, *non-kin*) pairs to the 5-folds– we randomly mismatched pairs in each fold until the number of negative and positive pairs are the same in each fold (*i.e.*, negative pairs are added at random until it makes up 50% of the respective fold). Thus, the total number of positive and negative labels are equivalent.

For this task we included each feature, metric learning approach, and deep learning model listed above. We then fine-tuned the pre-trained CNN models on the FIW dataset, which is described in detail in the next subsection. To compare features, we computed cosine similarity between each pair, which was then compared to a threshold to classify each pair as either *kin* or *non-kin*.

Verification accuracy (*i.e.*, average of 5-folds) and *receiver operating characteristic* (ROC) curves were used to evaluate.

5.2.2 Training Philosophy

For ResNet-22 + CF, we fine-tuned the Centerface model on our FIW data. Training was done using four Titan X GPUs with a batch size of 256. The learning rate was initially set to 0.01, then drops to 0.001 and 0.0001 at the 800 and 1200 iterations, respectively. Training was complete after 1,600 iterations. The weight decay was set to 0.0005. For SphereFace [65], the settings are similar to ResNet-22+CF (*i.e.*, same batchsize, learning rate, weight decay, and number of iterations), and with the angular margin set to 4.

5.2.3 Results

As listed Table 5, *siblings* pairs types tended to score the highest, followed by *parent-child* types, and then *grandparent-grandchild*. Thus, the wider the generational gap, the wider between appearances of faces.

SphereFace, which was fine-tuned on FIW, outperformed other benchmarks with an average accuracy of 69.18%, which is 1.31% and 2.11% better than ResNet-22+CF and mDML, respectively, which were top the scoring methods prior to the recent release of SphereFace. Also, out of the pretrained CNNs, VGG-Face scored 3.55% higher than ResNet-22, and both outperformed the low-level features (*i.e.*, LBP & SIFT). From such, encodings from VGG-Face were used as features for the metric learning and AE methods. Besides LMNN and DLML, which improved score by 0.15% and 1.30%, the other metric learning methods actually worsened the performance of the descriptors extracted from the pretrained VGG-Face CNN. This infers that faces encoded via VGG-Face are more discriminative when used *off-the-shelf* than when certain metrics are learned on top.

We show a significant boost in performance when fine-tuning CNNs on FIW data– all features from CNNs outperform the conventional shallow methods. The results show that the deep learning models better encode the complex representation needed to discriminate between *kin*/*non-kin* (see Fig. 6). An improvement to these benchmarks, perhaps via a deep network designed specifically for this task, is certainly a direction for future work.

5.3 Family Classification

Family classification is a *one-to-many* problem. The goal is to determine which family an unseen subject came from. In other words, a set of families with a missing member to the model is provided. Then, the missing (*i.e.*, unseen) members get classified as being from one of the families (*i.e.*, closed form, as we currently assume that all members at test time belongs to one of the families modeled during training). We next review some details for this task.

TABLE 6
Family classification accuracy scores (%) using 564 families.

Run ID	Network(s)	Acc.
Run-1	VGG-Face, fc_7 (4,096D)+*one-vs-rest* SVMs	3.04
Run-2	VGG-Face, replaced softmax (564D)+fine-tuned	10.42
Run-3	ResNet-22 + softmax (564D)	14.17
Run-4	SphereFace (564D)	13.86
Run-5	ResNet-22 + CF (512D) + softmax (564D)	**16.18**

5.3.1 *Experimental Setting*

Data from 564 families by leaving a different single member out in each fold for testing, while data from all the other members were used for training (*i.e.*, leave-one-out w.r.t. family members). Families with at least 5 members were used. Thus, the data was split into 5-folds with no family overlap between folds (*i.e.*, a minimum of 4 family members for training and 1 for testing). Each fold contained roughly 2,700 images– about that many faces used to test each split, while about the rest, about 12,800 faces, were used for training (*i.e.*, a total of 13,420 images).

5.3.2 *Training Philosophy*

VGG-Face and ResNet-22 CNNs were fine-tuned on FIW by replacing the loss layers of the pre-trained CNNs with a softmax loss to predict the 564 family classes. There were a few differences: VGG-16 used a fixed learning rate of 0.0001, a batch size of 128, and trained for 800 iterations on one Titan X GPU; ResNet-22 used the same batch size and number of iterations, but with a larger learning rate 0.001, which was fixed too. For ResNet-22 + CF and SphereFace, we followed the same training process used for verification.

5.3.3 *Results*

We report the accuracy scores for five runs (see Table 6). As shown, the top-1 accuracy for modeling *one-vs-rest* linear SVMs on top of deep VGG-Face features was just 3.04%. Then, by replacing the softmax layer to target the number of families (*i.e.*, 564), and fine-tuning on FIW, the top-1 accuracy was improved (*i.e.*, +7.38% to 10.42%). ResNet-22, also fine-tuned by replacing softmax layer, showed the second to highest accuracy with 14.17%, which outscored the top performing CNN on verification (*i.e.*, SphereFace). The top performance was obtained with the fine-tuned ResNet-22 using Centerface (CF) loss with 16.18%.

5.4 Proposed Semi-Supervised Clustering

To demonstrate the effectiveness of our semi-supervised model, we cluster FIW data using various amounts of *family-level* labels as side information. We simulate two settings for evaluation– all data and just unlabeled data– shown as bold and dotted lines, respectively (see Fig. 7).

5.4.1 *Experimental Setting*

We used $23,979$ faces from 996 family classes. Faces were encoded using a pre-trained VGG-Face (*i.e.*, fc_7). We varied the ratio of unlabeled data to side information across the horizontal axis up to 50% percent of labeled clusters, while the y-axis denotes the clustering performance on the rest of the unlabeled data by NMI. We compared to a pair-wise constrained clustering method, LCVQE [70], which is also a K-means-based constrained clustering method and transforms the partition level side information into 'must-link' and 'cannot-link' constraints. We used K-means as a baseline (*i.e.*, no side-information).

5.4.2 *Results*

Fig. 7 shows that more side information consistently boosts the performance of our method. Even on the unlabeled data, our method exceeds the K-means baseline, which further validates the effectiveness of using our method in semi-automatic labeling scenarios. For LCVQE, the pair-wise constraints make the cluster structure unpredictable, vulnerable to deviate from the true one, and, thus, perform worse than the baseline. This shows that imposing hard constraints on side information, like 'must-link' and 'cannot-link', may even damper results. On the contrary, our model leverages the side information to only only improve when more is added.

5.5 Transfer-Learning Experiment

To demonstrate that FIW generalizes well, we fine-tune the ResNet CNN model on the entire dataset and assess the model on a smaller, non-overlapping image collection. Specifically, we achieve state-of-the-art performance using a fine-tuned CNN to encode faces of the renown KinWild datasets (see Table 7). For KinWild I, we get a 4% increase in performance (*i.e.*, from 78.4% to 82.4%). For KinWild II, there is a 5.6% improvement to 86.6%.

Notice the significant boost in accuracy for KinWild I when comparing F-D to all other types, and especially M-D. Clearly, the small sample size of this dataset does not properly represent the data distribution of these pair types, while FIW has noticeably less variance between scores of parent-child types. Regardless of the high score of *Bar Ilan University* (BIU) for type F-D , our fine-tuned network performs better on all other types, in average accuracy, and while providing less variation in type-specific scores. Again, this variance is considered to be caused by a small sample size, as there is less variation in score for the parent-child types of FIW.

TABLE 7
Accuracy (%) for KinWild I & II. CNN fine-tuned on FIW top scorer.

	KinWild-I					*KinWild-II*				
Method	FD	FS	MD	MS	Avg.	FD	FS	MD	MS	Avg.
LBP [56]	72.8	79.5	71.7	68.1	73.0	70.8	78.4	69.0	73.2	72.9
SIFT [57]	73.9	81.4	76.4	71.1	75.7	72.2	78.8	82.2	79.6	78.2
NRML (LBP)	81.4	69.8	67.2	72.9	72.8	79.2	71.6	72.2	68.4	72.9
NRML (HOG)	83.7	74.6	71.6	80.0	77.5	80.8	72.8	74.8	70.4	74.7
BIU (LBP)	85.5	76.5	69.9	74.4	76.6	84.2	79.5	76.0	77.0	79.2
BIU (HOG)	**86.9**	76.5	70.6	79.8	78.4	87.5	80.8	79.8	75.6	81.0
VGG-Face [36]	72.0	77.6	78.3	80.6	77.1	68.8	74.4	76.6	74.6	73.6
ResNet + CF	78.0	**83.7**	**87.0**	**80.8**	**82.4**	**87.7**	**86.0**	**86.7**	**87.4**	**86.6**

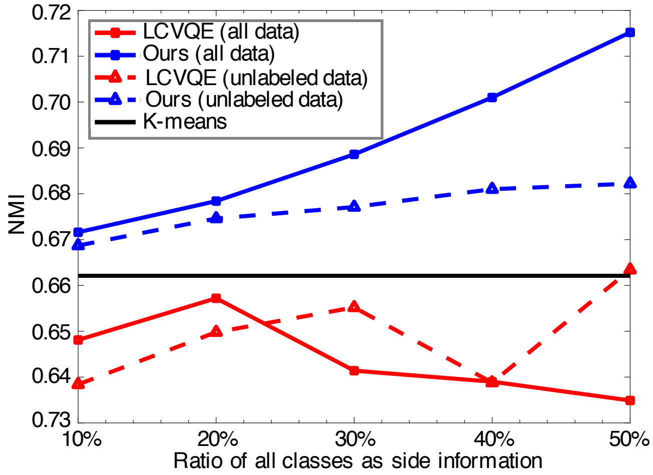


Fig. 7. Results for clustering families using different amounts of side information. As clearly depicted, our method obtains the top performance. Moreover, a distinct increase in NMI for our method is shown with an increase in the amounts of side information.

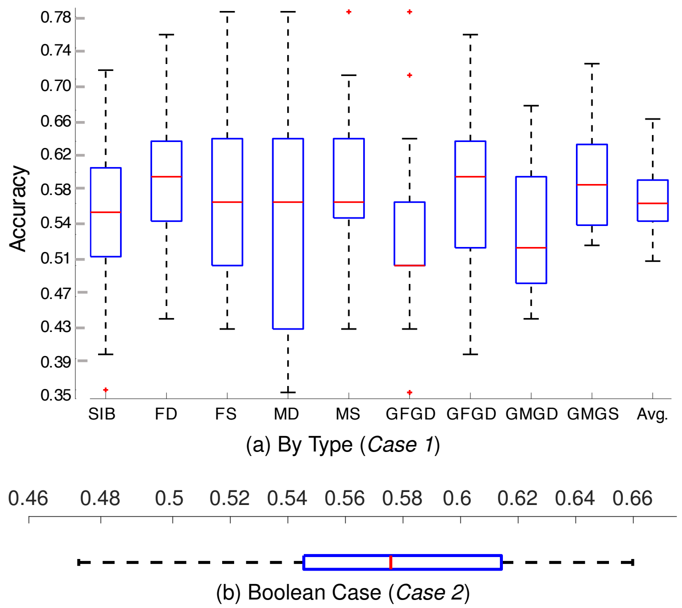
5.6 Human Performance on FIW

We evaluate human performance on kinship verification with a subset of FIW pairs. Although others conducted similar experiments [29], [30], [71], this was done with a larger sample set made up of more relationship types (*Case 1*). Additionally, an evaluation was done for the Boolean case only (*Case 2*). We now discuss experimental settings, results, and analyses of both human experiments.

5.6.1 Experimental Setting

First a list of pairs from FIW with a fair data distribution was sampled (*i.e.*, different and diverse families with faces of various ages). Faces for both positive and negative pairs were from different photos. Also, we used no more than one positive and negative sample per member. We rigorously examined and, in some ways, handcrafted the list to best control the experiment (*i.e.*, replaced face images of poorer quality and famous people). Thus, efforts were spent to better ensure a fair, unbiased assessment. We also only used faces to avoid evidence besides facial appearance influencing human responses [72]. The same list of images was used for both cases: evaluating pairs per specific relationship types and for the Boolean case only.

A Google Form was used to collect responses, and the university and social media networks to recruit volunteers. Answers were anonymous, although demographic information was collected (*i.e.*, ethnicity, country of origin, and gender). Some volunteers completed both experiments. However, scores and answers were not revealed. Also, there was nearly a year between when the two experiments were conducted, with the Boolean case being a follow-up experiment to analyze how specific relationship types influence responses. Users chose from predefined responses: *Related*, *Unrelated*, or *Skip*. Participants were asked to *Skip* if they had prior knowledge of one or both subjects, regardless of knowledge about the relationships (*i.e.*, skip any pair containing an identifiable face). Face pair-types were processed in no special order: a type-by-type basis for *Case 1*, then shuffled at random for *Case 2*. There was a total of 406 face pairs sampled from the 11 categories. Specifically, there were 50 for each *sibling* type (*i.e.*, 150 in total), 36 for each *parent-child* type (*i.e.*, 144 in total), and 28 for each *grandparent-grandchild* (*i.e.*, 112 in total).



(a) By Type (*Case 1*)

(b) Boolean Case (*Case 2*)

Fig. 8. Box plot for humans on kinship verification. *Case 1:* Relationship type dependent evalations. *Case 2:* Evaluations with type unspecified.

We had 75 and 110 volunteers for *Case 1* and 2, respectively. No training of any sort was provided. In both cases, the distribution of demographics was approximately 45% Caucasian, 35% Asian, 10% Hispanic/ Latino, 4% African American, and 1% Arab; 65% born in the United States, 30% from China, and 1-2% from South America, Middle East, and the Philippines; 55% males and 45% Females. No specific demographics were targeted (*i.e.*, a matter who volunteered on social media, per request of the authors, etc.). Future work could involve a greater emphasis on demographics, both in overall distribution of volunteers and intended analysis. Here, we hope to lay the framework for such a study, along with other interesting directions that assessing human ability to recognize kinship can take.

To compare human performance to benchmarks, we fine-tune SphereFace CNN on the 764 families that were not included in face pairs used for the human evaluation.

5.6.2 Results

We assess both human evaluations via box plots (see Fig. 8).

In *Case 1*, the minimum scores across most categories are below random (*i.e.*, $<$ %50). In response, we confirmed that no single person scored lowest in more than 1 of the 9 categories. Another observation is the distribution of averages, and its mean of 57.5%, had the smallest variance– no average below 50% or above 67.5%, which indicates that no single, or more than a few subjects, dominated the average scores for the better of the worse. Examining the pairs where errors were made, three conclusions can be made: (1) especially for relationship types spanning 1 or more generations (*i.e.*, parents and grandparents), the common

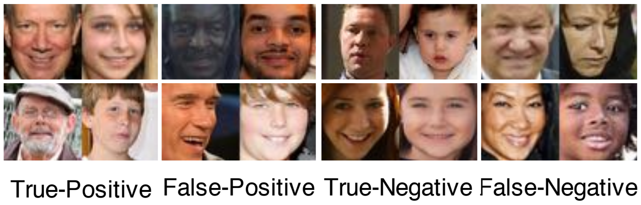


Fig. 9. Samples used for human evaluation. Each column displays pairs most commonly marked correctly and incorrectly, and in cases for where the correct answer were true and false. Each of these pairs were properly classified by the fine-tuned CNN.

pairs consistently marked incorrectly are cases were the face of the expected elder is at a younger age or the face of the descendant appears older (*e.g.*, grandfather in his thirties and grandson in his fifties); (2) different ethnic groups typically made common mistakes on face pairs of different backgrounds; (3) females often deviated from males on the mistakes made that are common and across different ethnic groups– varying females were always the top scorer, but never the same twice. Apparently, nature and nurture can play a role in humans' ability to do kinship verification as well. There are many interesting directions for future work (*e.g.*, even larger and more diverse subject pool, or samples with added semantics like full body views or entire photos with background context).

For *Case 2*, we evaluated humans' ability to recognize kinship in faces, but, this time, without specifying the relationship. From this, we were aiming to determine whether the relationship direction and face age impacted human responses. Overall, the mean values barely changed, however, the set of pairs commonly marked wrong did– relationship direction does seem to worsen human ability to recognize kinship when the direction of the relationship contradicts with the age appearance of face pairs; however, in cases without the age contradiction, knowledge of the relationship type helps humans to determine whether or not the face pairs are of that type (*i.e.*, even though the set of common pairs incorrectly classified changed, the overall mean did not, as the average fell between 57-58% in both cases). Fig. 9 shows face pairs most commonly classified correctly or incorrectly considering both cases.

Quantitatively, human performers scored an average of 57.5%. This is comparable to hand-crafted features such as LBP and SIFT, but nearly 15% lower than our fine-tuned CNN (*i.e.*, the SphereFace CNN fine-tuned for this experiment scores 72.15%).

6 Discussion and Future Work

The labels of FIW are dynamic in structure– labels can be parsed to use the data in various ways. For instance, siblings can be split between those who share one and both parents. Even a slight change in paradigm can drastically change the study– use both parents for verification (*i.e.*, tri-subject verification [12]); use child photos only to test with for family classification. Besides, we still need to improve our visual recognition capability for kinship in current benchmarks. Then, it only seems natural to aim for fine-grained categorization of entire family trees (*i.e.*, the ultimate achievement). On a different note, generative modeling is another interesting research track to pursue (*e.g.*, given a couple and predict the offspring, or samples of their baby and predict the baby's appearance as an adult). Even other pair types (*e.g.*, great- and great-great-grandparents, cousins, aunts, uncles, etc.). Also, the labeling framework introduced in this work could be used to add video data to the families of FIW, which can be served as a resource for template- based search and retrieval, or even consider emotional responses and facial expressions of family members.

We expect that as researchers advance this problem, FIW and its uses too will advance, and especially when considering the potential for interdisciplinary collaborations– Whether nature-based studies, generative or predictive modeling, or security-based. We hope FIW inspires new types of problems, and anticipate the list of uses to only grow when FIW is in the hands of researchers worldwide. In the end, the aim here is to attract more experts to the problem of kinship recognition.

7 Conclusion

Families In the Wild (FIW) is the first large-scale dataset available for visual kinship recognition. We annotated complex hierarchical relationships with only a small team in a fast and efficient manner– providing the largest labeled collection of family photos to-date. FIW was structured to support multiple tasks with its dynamic label structure. We provided several benchmarks for kinship verification and family classification. Pre-trained CNNs were used as *off-the-shelf* face encoders, which outperformed conventional methods. Results for both tasks were further improved by fine-tuning the CNN models on FIW. We measured human observers and compared their performance to the machine vision algorithms, showing that CNN models already surpass humans in recognizing kinship.

The size of FIW, along with the labels representing complex tree structures of $1,000$ families, makes it difficult to pinpoint the exact directions FIW will lead. Improving upon the benchmarks is one route, which is the focus of past, current, and future data challenges based on FIW. Also, additional task evaluations (*e.g.*, search & retrieval and tri-subject), along with cross-discipline studies (*e.g.*, nature-based and human perception) are also promising directions.

References

[1] R. Fang, K. D. Tang, N. Snavely, and T. Chen, "Towards computational models of kinship verification," in *17th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2010, pp. 1577–1580.

[2] X. Chen, L. An, S. Yang, and W. Wu, "Kinship verification in multilinear coherent spaces," *Multimedia Tools and Applications*, 2015.

[3] Y. Chen, H. Hu, S. Cao, and B. Ma, "Sparse coding based kinship recognition," in *Multimedia Technology IV: Proceedings of the 4th International Conference on Multimedia Technology, Sydney, Australia, 28-30 March 2015*. CRC Press, 2015, p. 115.

[4] A. Dehghan, E. 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*, 2014, pp. 1757–1764.

[5] X. Duan and Z.-H. Tan, "A feature subtraction method for image based kinship verification under uncontrolled environments," in *International Conference on Image Processing (ICIP)*. IEEE, 2015.

[6] X. Duan and Z. H. Tan, "Neighbors based discriminative feature difference learning for kinship verification," in *International Symposium on Visual Computing*. Springer, 2015, pp. 258–267.

[7] R. Fang, A. Gallagher, T. Chen, and A. Loui, "Kinship classification by modeling facial feature heredity," in *International Conference on Image Processing (ICIP)*. IEEE, 2013, pp. 2983–2987.

[8] Y. Guo, H. Dibeklioglu, and L. van der Maaten, "Graph-based kinship recognition," in *ICPR*, 2014, pp. 4287–4292.

[9] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *Computer Vision–ACCV 2014*. Springer, 2014, pp. 252–267.

[10] L. Kou, X. Zhou, M. Xu, and Y. Shang, "Learning a genetic measure for kinship verification using facial images," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[11] 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)*. IEEE, 2015, pp. 1–6.

[12] X. Qin, X. Tan, and S. Chen, "Tri-subject kinship verification: Understanding the core of a family," *CoRR*, vol. abs/1501.02555, 2015.

[13] T. F. Vieira, A. Bottino, and I. U. Islam, "Automatic verification of parent-child pairs from face images," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2013, pp. 326–333.

[14] T. F. Vieira, A. Bottino, A. Laurentini, and M. De Simone, "Detecting siblings in image pairs," *The Visual Computer*, vol. 30, no. 12, pp. 1333–1345, 2014.

[15] X. Wang and C. Kambhamettu, "Leveraging appearance and geometry for kinship verification," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 5017–5021.

[16] S. Xia, M. Shao, and Y. Fu, "Kinship verification through transfer learning," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. AAAI Press, 2011, pp. 2539–2544.

[17] M. Xu and Y. Shang, "Kinship verification using facial images by robust similarity learning," *Math Problems in Engineering*, 2016.

[18] H. Yan, J. Lu, W. Deng, and X. Zhou, "Discriminative multimetric learning for kinship verification," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 7, pp. 1169–1178, 2014.

[19] H. Yan, J. Lu, and X. Zhou, "Prototype-based discriminative feature learning for kinship verification," *IEEE Transactions on Cybernetics*, vol. 45, no. 11, pp. 2535–2545, 2015.

[20] L. Zhang, K. Ma, H. Nejati, L. Foo, T. Sim, and D. Guo, "A talking profile to distinguish identical twins," *Image and Vision Computing*, vol. 32, no. 10, pp. 771–778, 2014.

[21] 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*. ACM, 2011.

[22] X. Zhou, Y. Shang, H. Yan, and G. Guo, "Ensemble similarity learning for kinship verification from facial images in the wild," *Information Fusion*, vol. 32, pp. 40–48, 2016.

[23] 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 & Gesture Recognition (FG)*. IEEE, 2018.

[24] M. Shao, S. Xia, and Y. Fu, "Identity and kinship relations in group pictures," in *Human-Centered Social Media Analytics*. Springer, 2014, pp. 175–190.

[25] J. Zhang, S. Xia, M. Shao, and Y. Fu, "Family photo recognition via multiple instance learning," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 2017.

[26] S. Xia, M. Shao, and Y. Fu, "Toward kinship verification using visual attributes," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 549–552.

[27] 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 2016 ACM on Multimedia Conference*, ser. MM '16. New York, NY, USA: ACM, 2016, pp. 242–246. [Online]. Available: http://doi.acm.org/10.1145/2964284.2967219

[28] S. Xia, M. Shao, J. Luo, and Y. Fu, "Understanding kin relationships in a photo," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1046–1056, 2012.

[29] S. X. M. Shao and Y. Fu, "Genealogical face recognition based on ub kinface database," in *IEEE CVPR Workshop on Biometrics*, 2011.

[30] 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*, vol. 36, no. 2, pp. 331–345, 2014.

[31] J. Lu, J. Hu, V. E. Liong, X. Zhou, A. Bottino, I. Ul Islam, T. Figueiredo Vieira, X. Qin, X. Tan, S. Chen *et al.*, "The fg 2015 kinship verification in the wild evaluation," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–7.

[32] J. Lu, J. Hu, X. Zhou, J. Zhou, M. Castrillón-Santana, J. Lorenzo-Navarro, L. Kou, Y. Shang, A. Bottino, and T. Figuieiredo Vieira, "Kinship verification in the wild: The first kinship verification competition," in *IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2014, pp. 1–6.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[35] L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[36] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[37] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[38] K. Zhang, Y. Huang, C. Song, H. Wu, and L. Wang, "Kinship verification with deep convolutional neural networks," in *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, September 2015, pp. 148.1–148.12.

[39] A. Dehghan, E. G. Ortiz, R. Villegas, and M. Shah, "Who do i look like? determining parent-offspring resemblance via gated autoencoders," in *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2014.

[40] C. Xiong, L. Liu, X. Zhao, S. Yan, and T.-K. Kim, "Convolutional fusion network for face verification in the wild," 2015.

[41] S. Wang, J. P. Robinson, and Y. Fu, "Kinship verification on families in the wild with marginalized denoising metric learning," in *12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2017.

[42] X. Wu, E. Boutellaa, X. Feng, and A. Hadid, "Kinship verification from faces: Methods, databases and challenges," in *2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. IEEE, 2016, pp. 1–6.

[43] H. Liu and Y. Fu, "Clustering with partition level side information," in *Proceedings of International Conference on Data Mining*, 2015.

[44] C. W. Leong, R. Mihalcea, and S. Hassan, "Text mining for automatic image tagging," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, ser. COLING '10. PA, USA: Association for Computational Linguistics, 2010.

[45] E. Law, B. Settles, and T. Mitchell, "Learning to tag using noisy labels," in *Proc. ECML*, 2010, pp. 1–29.

[46] D. Wang, S. C. Hoi, Y. He, and J. Zhu, "Mining weakly labeled web facial images for search-based face annotation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 166–179, 2014.

[47] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[48] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[49] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[50] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[51] A. G. Bottino, M. De Simone, A. Laurentini, and T. Vieira, "A new problem in face image analysis: finding kinship clues for siblings pairs," 2012.

[52] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 363–370.

[53] H. Liu, Z. Tao, and Y. Fu, "Partition level constrained clustering," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[54] B. Mirkin, "Reinterpreting the category utility function," *Machine Learning*, 2001.

[55] J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen, "K-means-based consensus clustering: A unified view," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 1, pp. 155–169, 2015.

[56] 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*, vol. 28, no. 12, pp. 2037–2041, 2006.

[57] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 886–893 vol. 1.

[58] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[59] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 209–216.

[60] X. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16. MIT, 2004, p. 153.

[61] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[62] Y. Peng, S. Wang, and B.-L. Lu, "Marginalized denoising autoencoder via graph regularization for domain adaptation," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 156–163.

[63] Z. Ding, S. Suh, J.-J. Han, C. Choi, and Y. Fu, "Discriminative low-rank metric learning for face recognition," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–6.

[64] J. P. Robinson, M. Shao, H. Zhao, Y. Wu, T. Gillis, and Y. Fu, "Recognizing families in the wild (rfiw)," in *Proceedings of theh 2017 ACM Workshop on RFIW17*, 2017.

[65] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[66] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, 2004.

[67] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 529–534.

[68] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

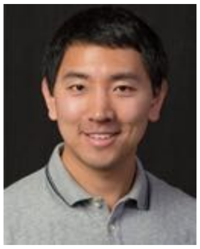
[69] J. P. Robinson and Y. Fu, "Pre-trained d-cnn models for detecting complex events in unconstrained videos," in *SPIE Commercial+ Scientific Sensing and Imaging*. International Society for Optics and Photonics, 2016, pp. 98 710O–98 710O.

[70] D. Pelleg and D.Baras, "K-means with large and noisy constraint sets," in *Proceedings of European Conference on Machine Learning*, 2007, pp. 674–682.

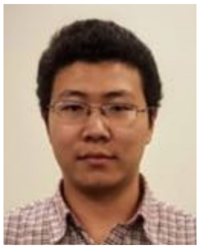
[71] M. F. Dal Martello and L. T. Maloney, "Where are kin recognition signals in the human face?" *Journal of Vision*, vol. 6, no. 12, 2006.

[72] L. Best-Rowden, S. Bisht, J. C. Klontz, and A. K. Jain, "Unconstrained face recognition: Establishing baseline human performance via crowdsourcing," in *Biometrics (IJCB), 2014 IEEE International Joint Conference on*. IEEE, 2014, pp. 1–8.

Joseph P. Robinson received BS in electrical and computer engineering ('14) and is pursuing a PhD in computer engineering at Northeastern University (NEU). His research is applied machine vision, with emphasis on faces, deep learning, multimedia, and large databases. He led efforts in TRECVid début. He led built several image & video databases– most notably FIW. Robinson has served throughout the community: organizer/ chair of several workshops (*e.g.*, RFIW@ACMMM, AMFG@CVPR, RFIW@FG, NECV), PC member (*e.g.*, FG, MIRP, MMEDIA), reviewer (*e.g.*, IEEE Transactions on– Biomedical Circuits and Systems, Image Processing, Pattern Analysis and Machine Intelligence), and leading positions like president of NEU IEEE and Relations Officer of IEEE SAC R1 Region. He did two NSF REUs ('10 & '11), co-op at Analogic Corporation & BBN Technology, and internships at MIT-LL ('14) & STR ('16 & '17).



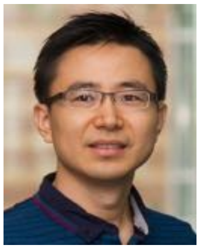
Ming Shao received the B.E. degree in computer science, the BS degree in applied mathematics, and the M.E. degree in computer science from Beihang University, Beijing, China, in 2006, 2007, and 2010, respectively. He received the Ph.D. degree in computer engineering from NEU, 2016. He is a tenure-track Assistant Professor affiliated with College of Engineering at University of Massachusetts Dartmouth since 2016 Fall. His current research interests include sparse modeling, low-rank matrix analysis, deep learning, and applied machine learning on social media analytics. He was the recipient of the Presidential Fellowship of State University of New York at Buffalo from 2010 to 2012, and the best paper award winner of IEEE ICDM 2011 Workshop on Large Scale Visual Analytics. He has served as the reviewers for IEEE journals: IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Knowledge and Data Engineering, etc.



Yue Wu Is a Ph.D student at NEU. He received the M.S and B.S. degree in Beijing University of Posts and Telecommunications at 2016 and 2013. His current research interests are face tracking, face alignment, face recognition, object detection and deep learning.

Hongfu Liu received his BS and MS in Management Information Systems from the School of Economics and Management, Beihang University, in 2011 and 2014 respectively. He is currently pursuing the PhD degree in NEU. His research interests generally focus on data mining and machine learning, with special interests in ensemble learning.

Timothy Gillis is a 3rd year undergraduate student studying computer science and computer engineering at NEU. His interests include computer vision and deep learning, mainly with respect to faces.



Yun Fu (S07-M08-SM11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xian Jiaotong University, China, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. He is an interdisciplinary faculty member affiliated with College of Engineering and the College of Computer and Information Science at NEU since 2012. His research interests are Machine Learning, Computational Intelligence, Big Data Mining, Computer Vision, Pattern Recognition, and Cyber-Physical Systems. He has extensive publications in leading journals, books/book chapters and international conferences/workshops. He serves as associate editor, chairs, PC member and reviewer of many top journals and international conferences/workshops. He received seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; seven Best Paper Awards from IEEE, IAPR, SPIE, SIAM; three major Industrial Research Awards from Google, Samsung, and Adobe, etc. He is currently an Associate Editor of the IEEE Transactions on Neural Networks and Leaning Systems (TNNLS). He is fellow of SPIE and IAPR, a Lifetime Senior Member of ACM, Lifetime Member of AAAI, OSA, and Institute of Mathematical Statistics, member of Global Young Academy (GYA), AAAS, INNS and Beckman Graduate Fellow during 2007-2008.