Frankenstein: Learning Deep Face Representations Using Small Data

Guosheng Hu, *Member, IEEE*, Xiaojiang Peng, Yongxin Yang, Timothy M. Hospedales, and Jakob Verbeek

***Abstract*— Deep convolutional neural networks have recently proven extremely effective for difficult face recognition problems in uncontrolled settings. To train such networks, very large training sets are needed with millions of labeled images. For some applications, such as near-infrared (NIR) face recognition, such large training data sets are not publicly available and difficult to collect. In this paper, we propose a method to generate very large training data sets of synthetic images by compositing real face images in a given data set. We show that this method enables to learn models from as few as 10 000 training images, which perform on par with models trained from 500 000 images. Using our approach, we also obtain state-of-the-art results on the CASIA NIR-VIS2.0 heterogeneous face recognition data set.**

***Index Terms*— Face recognition, deep learning, small training data.**

I. Introduction

In RECENT years, deep learning methods, and in particular convolutional neural networks (CNNs), have achieved considerable success in a range of computer vision applications including object recognition [25], object detection [10], semantic segmentation [37], action recognition [46], and face recognition [42]. The recent success of CNNs stems from the following facts: (i) big annotated training datasets are currently available for a variety of recognition problems to learn rich models with millions of free parameters; (ii) massively parallel GPU implementations greatly improve the training efficiency of CNNs; and (iii) new effective CNN architectures are being proposed, such as the VGG-16/19 networks [47], inception networks [55], and deep residual networks [13].

Manuscript received June 20, 2016; revised December 20, 2016 and May 25, 2017; accepted September 12, 2017. Date of publication September 26, 2017; date of current version November 3, 2017. This work was supported in part by European Unions Horizon 2020 Research and Innovation Program under Grant 640891, in part by the Science and Technology Plan Project of Hunan Province under Grant 2016TP1020, in part by the Natural Science Foundation of China under Grant 61502152, and in part by the French research agency contracts under Grant ANR-16-CE23-0006 and Grant ANR-11-LABX-0025-01. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gang Hua. *(Corresponding author: Xiaojiang Peng.)*

G. Hu and J. Verbeek are with CNRS, Grenoble INP, LJK, Université Grenoble Alpes, Inria, 38000 Grenoble, France (e-mail: guosheng.hu@inria.fr; jakob.verbeek@inria.fr).

X. Peng is with Hengyang Normal University, Hengyang 421008, China (e-mail: xiaojiangp@gmail.com).

Y. Yang is with Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: yongxin.yang@qmul.ac.uk).

T. M. Hospedales is with The University of Edinburgh, Edinburgh EH8 9JS, U.K. (e-mail: t.hospedales@ed.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2756450

Good features are essential for object recognition, including face recognition. Conventional features include linear functions of the raw pixel values, including Eigenface (Principal Component Analysis) [57], Fisherface (Linear Discriminant Analysis) [3], and Laplacianface (Locality Preserving Projection) [14]. Such linear features were later replaced by hand-crafted local non-linear features, such as Local Binary Patterns [1], Local Phase Quantisation (LPQ) [2], and Fisher vectors computed over dense SIFT descriptors [45]. Note that the latter is an example of a feature that also involves unsupervised learning. These traditional features achieve promising face recognition rates in constrained environments, as represented for example in the CMU PIE dataset [44]. However, using these features face recognition performance may degrade dramatically in uncontrolled environments, such as represented in the Labeled Faces in the Wild (LFW) benchmark [19]. To improve the performance in such challenging settings, metric learning can be used, see [5], [11], [59]. Metric learning methods learn a (linear) transformation of the features that pulls the objects that have the same label closer together, while pushing the objects that have different labels apart.

Although hand-crafted features and metric learning achieve promising performance for uncontrolled face recognition, it remains cumbersome to improve the design of hand-crafted local features (such as SIFT [28]) and their aggregation mechanisms (such as Fisher vectors [40]). This is because the experimental evaluation results of the features cannot be automatically fed back to improve the robustness to nuisance factors such as pose, illumination and expression. The major advantage of CNNs is that all processing layers, starting from the raw pixel-level input, have configurable parameters that can be learned from data. This obviates the need for manual feature design, and replaces it with supervised data-driven feature learning. Learning the large number of parameters in CNN models (millions of parameters are rather a rule than an exception) requires very large training datasets. For example, the CNNs which achieve state-of-the-art performance on the LFW benchmark are trained using datasets with millions of labeled faces: Facebook's DeepFace [56] and Google's FaceNet [42] were trained using 4 million and 200 million training samples, respectively.

For some recognition problems large supervised training datasets can be collected relatively easily. For example the CASIA Webface dataset of 500 000 face images was collected semi-automatically from IMDb [63]. However, in many other cases collecting large datasets may be costly, and possibly

problematic due to privacy regulation. For example, thermal infrared imaging is ideal for low-light nighttime and covert face recognition applications [24], but it is not possible to collect millions of labeled training images from the internet for the thermal infrared domain. The lack of large training datasets is an important bottleneck that prevents the use of deep learning methods in such cases, as the models will overfit dramatically when using small training datasets [18].

To address this issue, the use of big synthetic training datasets has been explored by a number of authors [20], [33], [38], [39]. There are two important advantages of using synthetic data (i) one can generate as many training samples as desired, and (ii) it allows explicit control over the nuisance factors. For instance, we can synthesize face images of all desired viewpoints, whereas data collected from the internet might be mostly limited to near frontal views. Data synthesis has successfully been applied to diverse recognition problems, including text recognition [20], scene understanding [33], and object detection [39]. Several recent works [9], [16], [17], [30], [67] proposed 3D-aided face synthesis techniques for facial landmark detection and face recognition in the wild.

Data augmentation is another technique that is commonly used to reduce the data scarcity problem [35], [47]. This is similar to data synthesis, but more limited in that existing training images are transformed without affecting the semantic class label, e.g. by applying cropping, rotation, scaling, etc.

The main contribution of this paper is a solution for training deep CNNs using small datasets. To achieve this, we propose a data synthesis technique to expand limited datasets to larger ones that are suitable to train powerful deep CNNs. Specifically, we synthesize images of a 'virtual' subject c by compositing automatically detected face parts (eyes, nose, mouth) of two existing subjects a and b in the dataset in a fixed pattern. Images for the new subject are generated by composing a nose from an image of subject a with a mouth of an image of subject b. This is motivated by the observation that face recognition consists in finding the differences in the appearance and constellation of face parts among people. For a dataset with an equal number of faces per person, this method can increase a dataset of n images to one with n^2 images when using only 2 face parts (we use 5 parts in practice). A dataset like LFW can thus be expanded from a little over 10 000 images to a dataset of 100 million images.

Unlike existing face synthesis methods which use 3D models [9], [16], [17], [30], [67], our method is a pure 2D method which is much easier to implement. In addition, our method works on different tasks from [9], [30], [67]. Specifically, the methods [9], [67] are used for facial landmark detection, while ours for face recognition. The approach [30] assumes a relatively large training data (500 000 images) already exists, while we assume the training data is very small (10 000 images).

We experimentally demonstrate that the synthesized large training datasets indeed significantly improve the generalization capacity of CNNs. In our experiments, we generate a training set of 1.5 million images using an initial labeled dataset of only 10 000 images. Using the synthetic data we improve the face verification rate from 78.97% to 95.77% on LFW. In addition, the proposed face synthesis is also used for NIR-VIS heterogeneous face recognition [32] and improve the rank-1 face identification rate from 17.41% to 85.05%. With the synthetic data, we achieve state-of-the-art performance on both (1) LFW under the "unrestricted, label-free outside data" protocol and (2) CASIA NIR-VIS 2.0 database under rank-1 face identification protocol.

II. Related Work

Our work relates to three research areas that we briefly review below: face recognition using deep learning methods (Section II-A), face data collection (Section II-B), and data augmentation and synthesis methods (Section II-C).

A. Face Recognition Using Deep Learning

Since face recognition is a special case of object recognition, good architectures for general object recognition may carry over to face recognition. Schroff *et al.* [42] explored networks that are based on that of Zeiler and Fergus [65] and inception networks [55]. DeepID3 [51] uses aspects of both inception networks and the very deep VGG network [47]. Parkhi *et al.* [34] use the very deep VGG network, while Yi *et al.* [63] use 3×3 filters but fewer layers. Hu *et al.* [15] use facial attribute information to improve the face recognition performance.

DeepFace [56] uses a 3D model for pose normalization, by which all the faces are rotated to the frontal pose. In this way, pose variations are removed from the faces. Then an 8-layer CNN is trained using four million pose-normalized images.

DeepID [53], DeepID2 [50], DeepID2+ [54] all train an ensemble of small CNNs. The input of one small CNN is an image patch cropped around a facial part (face, nose, mouth, etc.). The same idea is also used in [27]. DeepID uses only a classification-based loss to train the CNN, while DeepID2 includes an additional verification-based loss function. To further improve the performance, DeepID2+ adds losses to all the convolutional layers rather than the topmost layer only.

All the above methods train CNNs using large training datasets of 500 000 images or more. To the best of our knowledge, only [18] uses small datasets to train CNNs (only around 10 000 LFW images) and achieves significantly worse performance on the LFW benchmark: 87% vs 97% or higher in [42], [54], and [56]. Clearly, sufficiently large training datasets are extremely important for learning deep face representations.

B. Face Dataset Collection

Since big data is important for learning a deep face representation, several research groups have collected large datasets with 90 000 up to 2.6 million labeled face images [5], [31], [34], [52], [63]. To achieve this, they collect face images from the internet, by querying for specific websites such as IMDb or general search engines for celebrity names. This data collection process is detailed in [34] and [63].

Existing face data collection methods have, however, two main weaknesses. First, and most importantly, internet-based

collection of large face datasets is limited to visible spectrum images, and is not applicable to collect e.g. infrared face images. Second, the existing collection methods are expensive and time-consuming. It results from the fact that automatically collected face images are noisy, and manual filtering has to be performed to remove incorrectly labeled images [34].

The difficulty of collecting large datasets in some domains, e.g. for infrared imaging, motivates the work presented in this paper. To address this issue we propose a data synthesis method that we describe in the next section.

C. Data Augmentation and Synthesis

The availability of large supervised datasets is the key for machine learning to succeed, and this is true in particular for very powerful deep CNN models with millions of parameters. To alleviate data scarcity in visual recognition tasks, data augmentation has been used to add more examples by applying simple image transformations that do not affect the semantic-level image label, see e.g. [7]. Examples of such transformations are horizontal mirroring, cropping, small rotations, etc. Since it is not always clear in advance which (combinations of) transformations are the most effective to generate examples that improve the learning the most, Paulin *et al.* [35] proposed to learn which transformations to exploit.

Data augmentation, however, is limited to relatively simple image transformations. Out-of-plane rotations, for example, are hard to accomplish since they would require some degree of 3D scene understanding from a single image. Pose variations of articulated objects are another example of transformations that are non-trivial to obtain, and generally not used in data augmentation methods.

Training models from synthetic data can overcome such difficulties, provided that sufficiently accurate object models are available. Recent examples where visual recognition systems have been trained from synthetic data include the following. Shotton *et al.* [43] train randomized decision forests for human pose estimation from synthesized 3D depth data. Jaderberg *et al.* [20] use synthetic data to train CNN models for natural scene text recognition. Su *et al.* [48] use synthetic images of objects to learn a CNN for viewpoint estimation. Papon and Schoeler [33] train a multi-output CNN that predicts class, pose, and location of objects from realistic cluttered room scenes that are synthesized on the fly. Weinmann *et al.* [60] synthesize material images under different viewing and lighting conditions based on detailed surface geometry measurements, and use these to train a recognition system using a SIFT-VLAD representation [21]. Ronzantsev *et al.* [39] use rough 3D models to synthesize new views of real object category instances. They show that this outperforms more basic data augmentation using crops, flips, rotations, etc.

Data synthesis techniques are also used for face analysis. To improve the accuracy of facial landmark detection in the presence of large pose variations [9], [67], a 3D morphable face models is used to synthesize face images in arbitrary poses. Similar data synthesis techniques are also used for pose-robust face recognition [30]. Unlike 3D solutions, we propose a 2D data synthesis method to solve the problem of training deep CNNs using very limited training data.

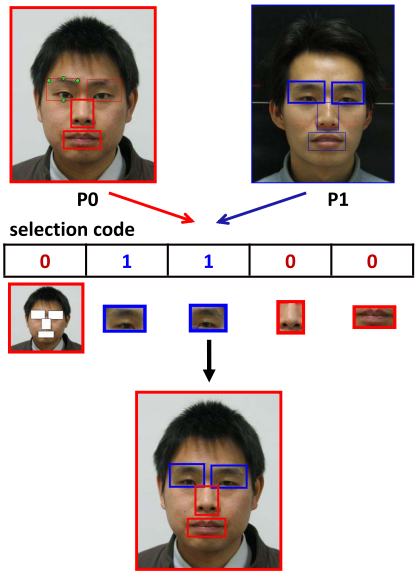

Fig. 1. Illustration of the face synthesis process using five parts: left-eye, right-eye, nose, mouth and the rest. Parent images P0 and P1 (top) are mixed by using the eyes of P1 and the other parts of P0 to form the synthetic image (bottom).

III. Synthetic Data Engine

Human faces are well structured in the sense that they are composed of parts (eyes, nose, mouth, etc.) which are organized in a relatively rigid constellation. Face recognition relies on differences among people in the appearance of facial parts and their constellation. Motivated by this, our synthetic face images are generated by swapping one or more facial parts among existing "parent" images. In our work we use five face parts: right eye (RE), left eye (LE), nose (N), mouth (M) and the rest (R). See Figure 1 for an illustration. For simplicity, we only consider the synthesis using only two parent images in this work. Our synthesis method can easily be extended, however, to the scenario of more than two parent images.

A. Compositing Face Images

Suppose that we have an original dataset and let $\mathcal{S}$ denote the set subjects in the dataset, and let n_i denote the number of images of subject $i \in \mathcal{S}$. To synthesize an image, we select a tuple (i, j, c, s, t) where $i \in \mathcal{S}, j \in \mathcal{S}$ correspond to two subjects that will be mixed, and $s \in \{1, \ldots, n_i\}$ and $t \in \{1, \ldots, n_j\}$ are indices of images of i and j that will be used. The bitcode $c \in \{0, 1\}^5$ defines which parts will be taken from each subject. A zero at a certain position in c means that the corresponding part will be taken from i, otherwise it will be taken from j. There are only $2^5 - 2 = 30$ valid options for b, since the codes 00000 and 11111 correspond to the original images of s and t respectively, instead of synthetic ones.

To synthesize a new image, we designate one of parent images as the "base" image from which we use the R (the rest) part, and the other as the "injection" image from which

one or more parts will be pasted on the base image. Since the size of the facial parts in the two parent images are in general different, we re-size the facial parts of the injection image to that of the base image. The main challenge to implement the proposed synthesis method is to accurately locate the facial parts. Recently, many efficient and accurate landmark detectors have been proposed. We use four landmarks detected by the method of Zhang *et al.* [66] to define the rectangular region that corresponds to each face part.

We refer to each choice of (i, j, c) with $i \neq j$ as a "virtual subject" which consists of a mix of two existing subjects in the dataset. In total we can generate $30|\mathcal{S}|(|\mathcal{S}|-1)/2$ different virtual subjects, and for each of these we can generate $n_i \times n_j$ samples. Note that if we set $i = j$ we can in the same manner synthesize 30 n_i $(n_i-1)/2$ new images for an existing subject.

Although some works [30], [56] empirically verified the effectiveness of synthetic data, they did not give much insight into how. In our work, the synthetic data captures a dataset of richer intra-personal variations by generating a large number of images of the same identities, leading to a 'deeper' training set. Also, our engine can synthesize a large number of faces of new identities, generating a 'wider' training set. Thus the synthetic identities interpolate the whole space of pixel-identity mappings. Not surprisingly, a better CNN model can be trained using this deeper and wider training set. The methods of generating our deeper and wider training set are detailed in Section V-A.

B. Compositing Artefacts

The synthetic faces present two types of artefacts: (I) hard boundary effects, and (II) inconsistent/unnatural intra-personal variations (lighting, pose, etc.) between facial patches. These are illustrated in Fig. 2. Note that the type I artefacts are generated by not only our method but also 3D synthesis methods such as [12], [30], and [57]. As shown in the top-right side of Fig. 2, the artefacts created by 3D methods are due to inaccurate 3D model to 2D image fitting. The inaccurate fitting makes the synthetic faces extract the pixels from background rather than facial areas, leading to bad facial boundaries.

Despite the existence of these artefacts, this synthetic data is still useful for training strong face recognition models. This can be understood from several perspectives: (1) Type I artefacts are common to all the synthetic faces in the training set, therefore the CNN does not learn to rely on artefacts as discriminative features coding for identity, i.e., it learns artefact invariance. This means its performance is not compromised when subsequently presented with artefact-free images at testing-time. Other studies have also shown that synthetic data still improves recognition performance, despite the presence of type I artefacts [30]. (2) The artefacts can be regarded as noise, which has been shown to improve model generalisation in a variety of settings by increasing robustness and reducing overfitting. For example in the case of CNNs, training with data augmentation in the form of specifically designed deformation noise is important to obtain good recognition performance [64]; and in the case of de-noising auto encoders, training on images with noise, corruption and artefacts has been shown to improve face classification performance [58]. (3) As a concrete example to understand how type II artefacts can improve performance, consider two synthetic images with the same identity label, but one a type II artefact on the mouth caused by illumination, e.g., in Fig. 2 (bottom left). Training to predict these images as having the same identity means that the CNN learns an illumination-invariant feature for the mouth. And similarly for other intra-personal variations (such as pose, expression). Thus while some artefact images look strange, they are actually a powerful form of data augmentation that helps the CNN to learn robustness to all these nuisance factors.

Fig. 2. Top row: Type I hard boundary artefacts generated by our method (left) and 3D synthesis methods [12], [30], [56] (right). Bottom row: Type II artefacts due to inconsistencies in illumination (left) and pose (right) generated by our method. For our method we show the two parent images followed by the composited image.

IV. Face Recognition Pipeline

In this section we describe the different elements of our pipeline for face identification and verification in detail.

A. CNN Architectures

Face recognition in the wild is a challenging task. As described in Section II-A, the existing deep learning methods highly depend on big training data. Very little research investigates training CNNs using small data. Recently, Hu *et al.* [18] evaluated CNNs trained using small datasets. Due to the limited training samples, they found the performance of CNNs to be worse than handcrafted features such as high-dimensional features [6] (0.8763 vs 0.9318). In this work, we use a limited training set of around 10 000 images to synthesize a much larger one of around 1.5 million images for CNN training. The synthesized training data captures various deformable facial patterns that are important to improve the generalization capacity of CNNs.

We use two CNN architectures. The first one, from [18] has fewer filters and is referred as CNN-S. The second, from [63], is much larger and referred as CNN-L. These two architectures are detailed in Table I. Using the CNN-L model we achieve state-of-the-art performance on the LFW dataset [19] under 'unrestricted, label-free outside data' protocol.

B. NIR-VIS Heterogeneous Face Recognition

NIR-VIS (near-infrared to visual) face recognition is important in applications where probe images are captured by NIR

TABLE I
OUR TWO CNN ARCHITECTURES

CNN-L	CNN-S
conv1	
32×3×3, st.1; 64 × 3 × 3, st.1 x2 maxpool, st.2	16×3×3, st.1; 16 × 3 × 3, st.1 x2 maxpool, st.2
conv2	
64×3×3, st.1; 128 × 3 × 3, st.1 x2 maxpool, st.2	32×3×3, st.1 x2 maxpool, st.2
conv3	
96×3×3, st.1 ; 192 × 3 × 3, st.1 x2 maxpool, st.2	48×3×3, st.1 x2 maxpool, st.2
conv4	
128×3×3, st.1 ; 256 × 3 × 3, st.1 x2 maxpool, st.2	-
conv5	
160×3×3, st.1 ; 320 × 3 × 3, st.1 x7 avgpool, st.1	-
fully connected	
Softmax-5000	FC-160 Softmax-5000

cameras that use active lighting which is invisible to the human eye [32]. Gallery images are, however, generally only available in the visible spectrum. The existing methods for NIR-VIS face recognition include three steps: (i) illumination pre-processing, (ii) feature extraction, and (iii) metric learning. First, the NIR-VIS illumination differences cause the main difficulty of NIR-VIS face recognition. Therefore, illumination normalization methods are usually used to reduce these differences. Second, to reduce the heterogeneities of NIR and VIS images, illumination-robust features such as LBP are usually extracted. Third, metric learning is widely utilized, aiming at removing the differences of modalities and meanwhile keeping the discriminative information of the extracted features.

In this work, we also follow these three steps that are detailed in Section V-D. Unlike the existing work that extracts handcrafted features, we learn face representations using two CNN architectures described above. To our knowledge, we are the first to use deep CNNs for NIR-VIS face recognition. The main difficulty of training CNNs results from the lack of NIR training images, which we address via data synthesis.

C. Network Fusion

Fusion of multiple networks is a widely used strategy to improve the performance of deep CNN models. For example, in [47] an ensemble of seven networks is used to improve the object recognition performance due to complementarity of the models trained at different scales. Network fusion is also successfully applied to learn face representations. DeepID and its variants [50], [53], [54] train multiple CNNs using image patches extracted from different facial parts.

The heterogeneity of NIR and VIS images is intrinsically caused by the different spectral bands from which they are acquired. The images in both modalities, however, are reflective in nature and affected by illumination variations. Illumination normalization can be used to reduce such variability, at the risk of loosing identity-specific characteristics. In this work, we fuse two networks that are trained using the original and illumination-normalized images respectively. This network fusion significantly boosts the recognition rate.

D. Metric Learning

The goal of metric learning is to make different classes more separated, and instances in the same class closer. Most approaches learn a Mahalanobis metric

$$d_A^2(x_i, x_j) = (x_i - x_j)^T A(x_i - x_j), \tag{1}$$

which maximizes inter-class discrepancy, while minimizing intra-class discrepancy. Other methods, instead learn a generalized dot-product of the form

$$d_B^2(x_i, x_j) = x_i^T B x_j. \tag{2}$$

Metric learning methods are widely used for face identification and verification. Because identification and verification are two different tasks, different loss functions should be optimized to learn the metric. Joint Bayesian metric learning (JB) [5] and Fisher linear discriminant analysis (LDA) are probably the two most widely used metric learning methods for face verification and identification respectively. In particular, LDA can be seen as a method to learn a metric of the form of Eq. (1), while JB learns a verification function that can be written as a weighted sum of Eq. (1) and (2). In our work we use JB and LDA to improve the performance of face verification and identification respectively.

V. EXPERIMENTS

A. Data Synthesis Methods

Given a set of face images and their IDs, we define three strategies for synthesis: *Inter-Synthesis*, *Intra-Synthesis*, and *Self-Synthesis*. *Inter-Synthesis* synthesizes a new image using two parents from different IDs as shown in Fig. 1. The facial components of an *Intra-Synthesized* face are from different images with the same ID. *Self-synthesis* is a special case of *Intra-Synthesis*. Specifically, one given image synthesizes new images by swapping facial components of itself and its mirrored version. By virtue of *Self-Synthesis*, one input image can become maximum 32 images which have complementary information. In the view of NIR-VIS cross-modality, we also define 'cross-modality synthesis' which uses images from different modalities to synthesize a new one. Some synthetic images from the CASIA NIR-VIS 2.0 dataset with LSSF [61] illumination normalization are shown in Fig. 3. The reasons of using LSSF illumination normalization are detailed in Section V-D. As shown in Fig. 3, the results of *Intra-Synthesis* method are usually more natural than *Inter-Synthesis* method since the *Intra-Synthesis* method uses the same ID. However, as shown in the right of Fig. 3, *Intra-Synthesis* can also create artefacts due to large pose variations.

B. Implementation Details

Before face synthesis, all the raw images are aligned and cropped to size 100 × 100 as in [63] on both datasets. We train our models using images only from LFW and CASIA NIR-VIS2.0 databases. For the CNN-S model on both datasets, we set the learning rate as 0.001, and decrease it by 10 times every 4000 iterations, and stop training after 10K iterations. We find that dropout is not helpful for the small network, and

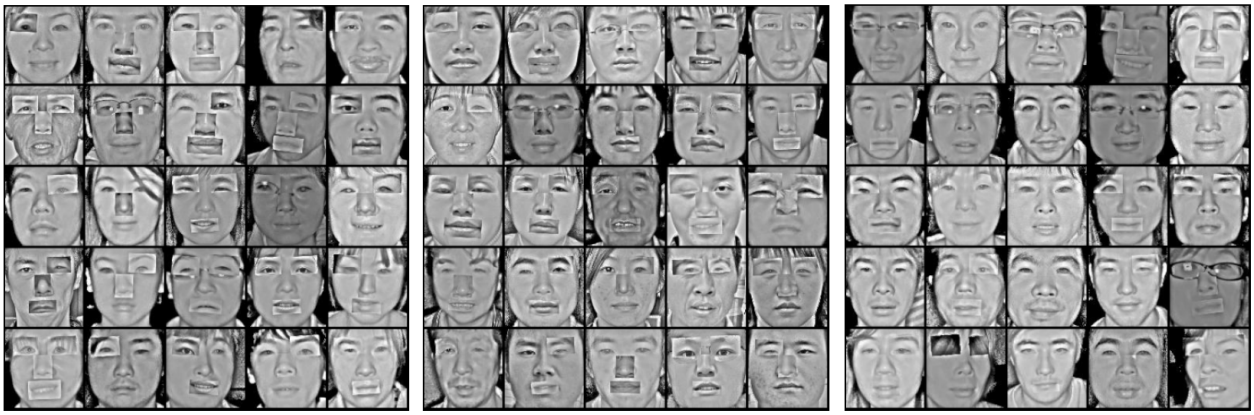

Fig. 3. Left: Inter-Synthesis. Middle: cross-modality Intra-Synthesis. Right: Intra-Synthesis.

TABLE II
TRAINING DATA SYNTHESIZED FROM LFW

		IDs	Images	Images/ID
Synthetic	Intra-Syn	5K	500K	100
	Inter-Syn	5K	1M	200
	Total	10K	1.5M	150
Raw		5K	10K	2

report results obtained without it. For the CNN-L model on the NIR-VIS dataset, we set the learning rate as 0.01, and decrease it by 10 times every 8000 iterations, and stop training after 20K iterations. For the CNN-L model on the LFW dataset, we set the learning rate as 0.01, and decrease it by 10 times every 120K iterations, and stop training after 200K iterations. We set dropout rate as 0.4 for the pool5 layer of the CNN-L model. For both CNN-S and CNN-L models, the batch size is 128, momentum is 0.9, and decay is 0.0005. Softmax loss function is used to guide CNN training. The features used in our recognition experiments with CNN-S and CNN-L are FC-160 (160D) and Pool5 (320D), respectively.

C. Face Recognition in the Wild

1) Database and Protocol: Labeled Faces in the Wild (LFW) [19] is a public dataset for unconstrained face recognition study. It contains 5,749 unique identities and 13 233 face photographs. The training and test sets are pre-defined in [19]. For evaluation, the full dataset is divided into ten splits, and each time nine of them are used for training and the left one for testing. Our work falls in the protocol of 'Unrestricted, Label-Free Outside Data' as we use the identity information to train the neural network (softmax loss). Meanwhile, all face images are aligned using a model trained on unlabeled outside data. As a benchmark for comparison, we report the mean and standard deviation of classification accuracy.

Under LFW protocol, the training set in each fold is different. Therefore, the size of synthetic data and the original raw LFW data in Table II is averaged over 10 folds. We generate 1.5 million training images, including 1 million 'Inter-Syn' ones and 0.5 million 'Intra-Syn' ones, as defined in Section V-A.

2) Analysis of CNN Model and Synthetic Data: We here analyse the trained model by visualising the synthetic images

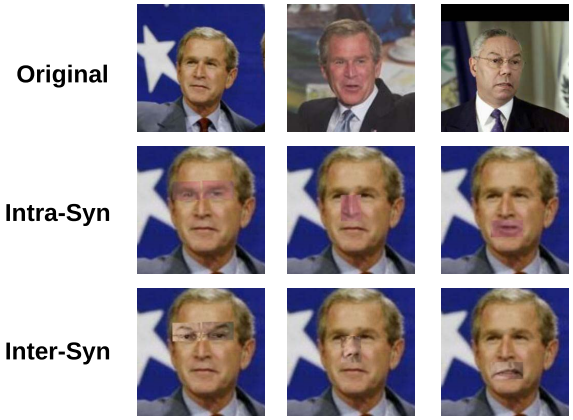

Fig. 4. Samples of Inter-Synthesis and Intra-Synthesis.

in feature space. We choose the images of two subjects (George W. Bush and Colin Powell), which have the largest number of images in LFW, and the synthetic images derived from the two subjects for analysis. To analyse the effects of replacing different facial components (eyes, nose, mouth), we only replace one patches of a particular facial component of Bush with ones from himself (Intra-Synthesis) or from Powell (Inter-Synthesis), as shown in Fig. 4. For each case, 100 images are synthesised. Therefore, there are 8 groups of images: 2 groups of original images (Bush and Powell), 3 Intra-Synthesis (one of three components is replaced by images of Bush) and 3 Inter-Synthesis (by images of Powell). These images are fed into one CNN-L to extract features, which are then projected to a PCA space. The first two PCA components of each feature are shown in Fig. 5.

Fig. 5(a)-(c) show the face distributions if one particular facial component is replaced. In Fig. 5(a), 3 identities, 'GeorgeWBush+intra-Eyes', 'ColinPowell' and 'inter-Eyes' are well separated. It means that the identity information is kept if Bush's eyes are replaced by those from himself. In contrast, a new identity space is generated by the synthetic images that replace Bush's eyes with Powell's. The same conclusion can be drawn from nose in Fig. 5(b). In Fig. 5(c), however, 'intra-Mouth' and 'inter-Mouth' are not well separated. It means that the mouth component is not very discriminative between people. Fig. 5(d) and (e) contrast the results when training with original and synthetic (Fig. 5(d))

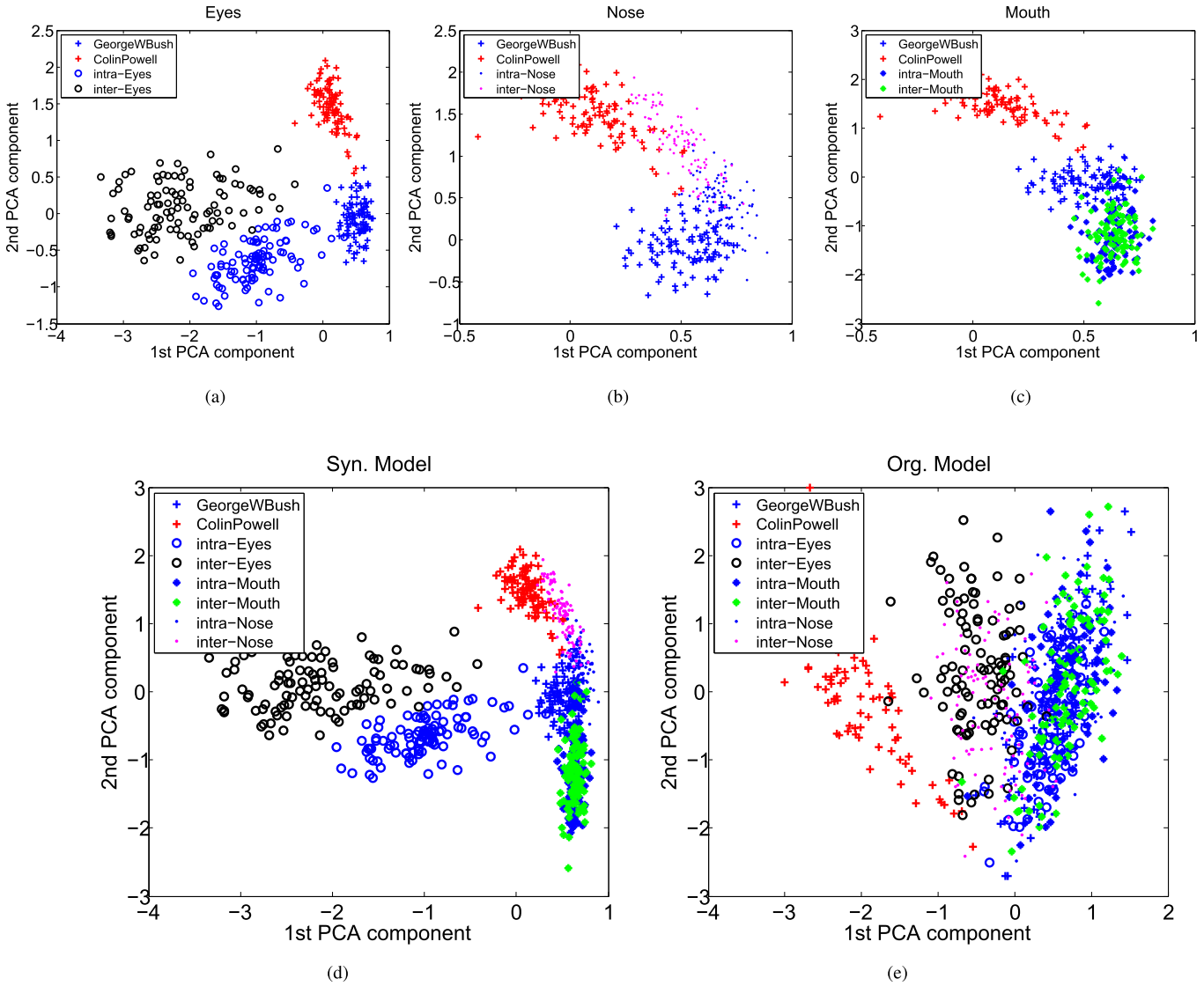

Fig. 5. Face distributions. 'GeorgeWBush' and 'ColinPowell' denote the original images from these 2 subjects. 'intra-X' and 'inter-X' denote the synthetic images with component X replaced by another example from Bush himself and from Powell respectively, as in Fig. 4. Features of (a)-(d) are extracted using a CNN trained using original and synthetic data, while (e) using original data only. One colour represents one (real or synthetic) subject.

versus original data only (Fig. 5(e)). Fig. 5(d) is relatively more discriminative for identity, particularly the synthetic identities. Thus we can see that training with synthetic data is important to interpolate the identity space, and thus achieve good results for unseen identities – as required at testing time.

3) Impact of Synthetic Data: Table III analyzes the importance of using the synthetic data. First, CNN-S trained using synthetic data ('Intra-Syn' and 'Inter-Syn') outperforms greatly the model trained using just the original LFW images, showing the importance of data synthesis. Second, 'Inter-Syn' works slightly better than 'Intra-Syn', since 'Inter-Syn' can capture richer facial variations. Third, combining 'Inter-Syn' and 'Intra-Syn' works better than either of them because they capture complementary variations. Fourth, averaging the features of 32 'Self-Syn' ('32-Avg' in Table III and defined in Section V-A) images works consistently better than that of one single test image ('single' in Table III). Fifth, CNN-L works consistently better than CNN-S using either original LFW or synthetic images because deeper architecture has stronger generalization capacity. Finally, metric learning further enhances the face recognition performance.

4) Impact of Image Blending: In Section III, we speculated the existence of artifacts ('hard boundaries') can improve the robustness of the model. We now experimentally investigate this by reducing such 'hard boundaries' using image blending. In particular, we implemented Poisson image editing [36] to smooth these boundaries. In Fig. 6, we show some results of Poisson image blending. From Fig. 6, as expected, Poisson blending does make the boundaries much smoother compared with our synthesis method. In Table III, we compare the results with and without Poisson image editing. We can see that recognition accuracy based on training data synthesised with 'hard' boundaries is, somewhat, higher than with Poisson blending. This supports the idea that 'hard' boundaries provide a source of noise that is beneficial in making network training more robust detailed in Section III.

TABLE III
COMPARISON OF SYNTHETIC DATA METHODS ON LFW

Architecture	Metric learning	Training data	single (%)	32-Avg (%)
CNN-S	-	Original	78.97 ± 0.78	-
		Intra-Syn+Original	83.03 ± 0.56	83.93 ± 0.49
		Inter-Syn+Original	83.18 ± 0.74	84.35 ± 0.65
		Intra-Syn+Inter-Syn+Original	85.61 ± 0.71	86.98 ± 0.57
CNN-L	-	Original	85.03 ± 0.98	-
	JB [5]	Original	87.03 ± 0.69	-
	-	Intra-Syn+Inter-Syn+Original	94.88 ± 0.66	95.13 ± 0.53
	JB [5]	**Intra-Syn+Inter-Syn+Original**	$\mathbf{95.32 \pm 0.38}$	$\mathbf{95.77 \pm 0.38}$
	-	Intra-Syn+Inter-Syn+Original (blending)	94.27 ± 0.65	94.46 ± 0.51
	JB [5]	Intra-Syn+Inter-Syn+Original (blending)	94.61 ± 0.35	95.05 ± 0.34

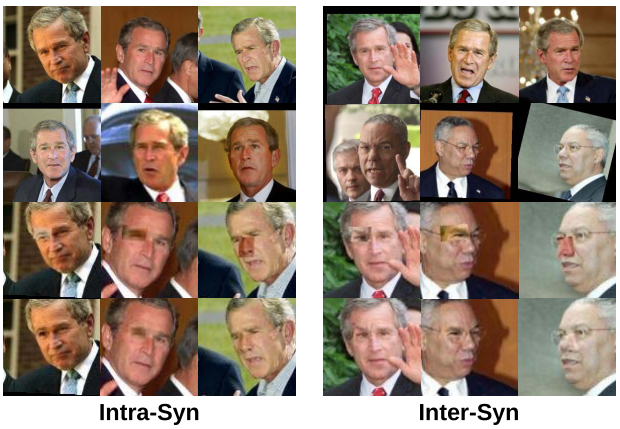

Fig. 6. Comparison of synthesis with/without blending. Row 1-2: Input image pairs, Row 3: Our synthesis with 'hard boundaries', Row 4: Poisson blending.

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON LFW

Methods	Accuracy (%)
Fisher vector faces [45]	93.03 ± 1.05
HPEN [68]	95.25 ± 0.36
MDML-DCPs [8]	95.58 ± 0.34
The proposed	$\mathbf{95.77 \pm 0.38}$

5) Comparison With the State-of-the-Art: Table IV compares our method with state-of-the-art methods. All methods listed in Table IV except ours use hand-crafted features, again underlining the difficulty of training deep CNNs with small data. The best deep learning solution [49] recorded in official benchmark achieves 91.75%, and ours is 4% better. In addition, most of state-of-the-art solutions rely on extremely high dimensional feature vectors derived from densely sampled local features on the face image. In contrast, we just use a 320-dimensional feature vector, which is much more compact.

6) Non-CNN Methods Using Synthetic Data: Above we demonstrated the effectiveness of synthetic data to train CNNs. We now consider its usefulness to improving methods based on traditional hand-crafted features. One typical hand-crafted feature used for unconstrained face recognition problem is high dimensional LBP (HD-LBP) [6]. We extract the HD-LBP feature using the open source code [4]. From Table V, we see that HD-LBP with JB metric learning trained using original LFW images works much better that without JB (89.02% vs 84.13%), showing the expected effectiveness of metric learning. More interestingly, we see that training JB using both original and synthetic images outperforms that using original images only, 91.03% vs 89.02%. This shows that our synthetic data approach is also useful in combination with conventional hand-crafted features.

TABLE V
HAND-CRAFTED FEATURES ON LFW

Methods	Accuracy (%)
HD-LBP	84.13 ± 1.76
HD-LBP+JB (original)	89.02 ± 1.11
HD-LBP+JB (original + synthetic)	91.03 ± 1.06

7) Impact of Synthetic Images on Larger Training Sets: Our original motivation was to learn effective face representations from small datasets. We demonstrated the effectiveness of generating synthetic data to expand a small training set (LFW, 5K identities and 10K images) in Table III. Recently, some bigger training sets of face images in the wild have been released, such as CASIA WEBFACE [63] (10K identities and 0.5M images). We conduct experiments using the latter dataset to assess whether our data synthesis strategy is also useful for such larger datasets. As before, we keep the ratio 2:1 of 'Inter-Syn' and 'Intra-Syn' synthetic images. To investigate the effects of the size of synthetic data, we generate six sets of synthetic images: {0.5M, 1M, 1.5M, 2M, 2.5M, 3M} images. We trained the CNN-L network using the original CASIA WEBFACE images, plus a variable amount of synthetic data. From Fig 7, we can see that the recognition rate on LFW increases with the amount of synthetic data. This demonstrates that our data synthesis strategy is still very effective even with relatively large datasets.

D. NIR-VIS Face Recognition

1) Database and Protocol: The largest face database across NIR and VIS spectrum so far is the CASIA NIR-VIS 2.0 face database [26]. It contains 17 580 images of 725 subjects which exhibit intra-personal variations such as pose and expression. This database includes two views: view 1 for parameter tuning and view 2 including 10 folds for performance evaluation. During test, the gallery and probe images are VIS and NIR images respectively, simulating the scenario of face recognition in the dark environment. The rank 1 identification

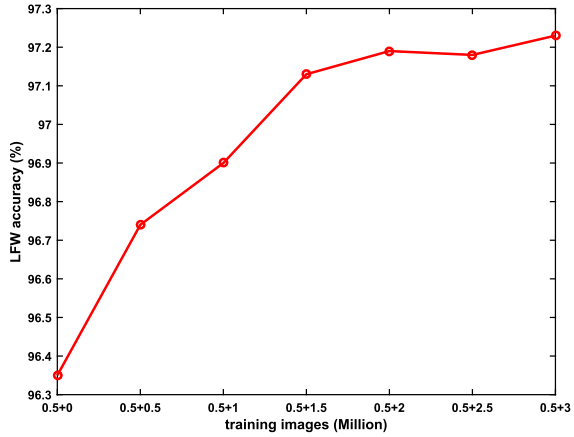

Fig. 7. Recognition rate (%) on LFW using the original CASIA WEBFACE training data (0.5M) and synthetic data (0-3M).

TABLE VI
SYNTHETIC DATA USING CASIA NIR-VIS2.0 DATABASE

		IDs	Images	Images/ID
Synthetic	Intra-Syn	357	90K	250
	Inter-Syn	1K	150K	150
	Total	1.4K	240K	170
Original		357	8.5K	23

rate including the mean accuracy and standard deviation of 10 folds are reported.

Because the images of CASIA NIR-VIS2.0 are from two modalities (NIR and VIS), we applied 'cross-modality synthesis' to synthesize new images. The size of synthesized data is detailed in Table VI.

2) Illumination Normalization and Feature Extraction: Illumination Normalization (IN) methods are usually used to narrow the gap between NIR and VIS images. To investigate the impact of IN, we preprocessed images using three popular IN methods: illumination normalization based on large-and small-scale features (LSSF) [61], Diffence-of-Gaussian filtering-based normalization (DOG), and single-scale retinex (SSR) [22]. We train CNN-S and CNN-L using illumination normalized and non-normalized images. For simplicity, only the images from CASIA NIR-VIS2.0 excluding synthetic ones are used. Fig. 8 shows the face recognition rates at different training iterations using different input images for the CNN-S and CNN-L networks. The results show the effectiveness of IN, and LSSF achieves the best performance due to its strong capacity of removing illumination effects without affecting identity information. As for the LFW experiments in Section V-C, CNN-L works better than CNN-S.

We also extracted LBP features from the LSSF-normalized images, and achieve 12.48 ± 3.1 in comparison with 17.41 ± 3.76 obtained with features from the CNN-L network. Showing again the superior performance of CNN learned features.

3) Effects of Synthetic Data: In practice, we find two problems with the synthetic data generated from the CASIA NIR-VIS2.0 dataset: (1) It cannot capture enough facial variations because it only has 357 subjects as shown in Table VI. (2) There are much fewer VIS images than

TABLE VII
EVALUATION OF THE IMPACT OF SYNTHETIC DATA

	Training Data		Accuracy(%)
	CASIA NIR-VIS2.0	LFW	
Baseline	Raw	-	17.41 ± 3.76
Synthetic Data	Raw+Syn	-	34.13 ± 2.13
	-	Raw	38.45 ± 2.08
	-	Raw+Syn	66.37 ± 1.45
	Raw+Syn	Raw+Syn	$\mathbf{68.97 \pm 1.24}$

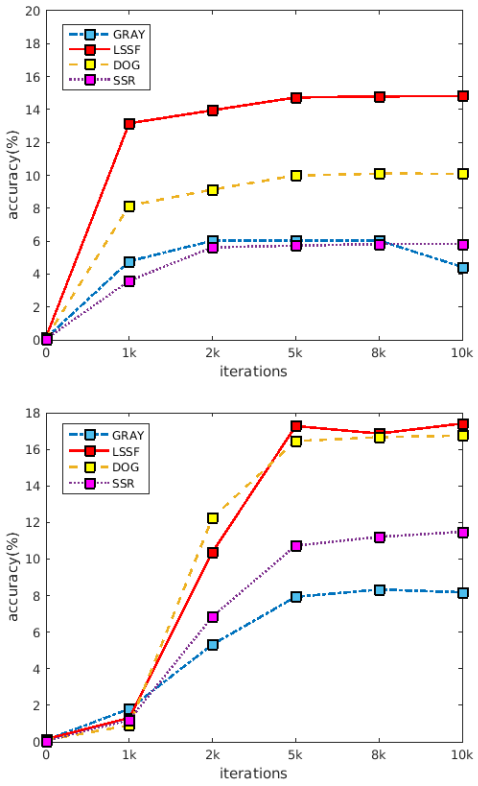

Fig. 8. Effect of illumination normalization methods using the CNN-S (top) and CNN-L networks.

NIR ones. To solve these two problems, we also use the synthetic data generated from LFW images defined in Table II.

Table VII compares the results achieved by these two sources of synthetic data. First, the accuracy achieved by using the synthetic data generated from CASIA NIR-VIS2.0 database is 34.13 ± 2.13, in comparison with 17.41 ± 3.76 without synthetic data. The significant improvement shows the effectiveness of data synthesis. Second, the model trained using raw and synthetic LSSF-normalized LFW images greatly outperforms those synthetic CASIA NIR-VIS2.0 images: 66.37 ± 1.45 vs. 38.45 ± 2.08, although NIR images are completely unseen during training. The reasons are 2-fold: (1) LFW images contains more subjects which can capture more facial variations as analyzed above. (2) LSSF can greatly reduce the gap between NIR and VIS, therefore, LSSF-normalized LFW synthetic images can generalize well to LSSF-normalized NIR images. To further improve the face recognition performance, we trained the network using both raw and synthetic data from both sources (LFW+CASIA NIR-VIS). The face recognition rate is improved from

TABLE VIII

COMPARISONS WITH THE STATE OF THE ART ON THE CASIA NIR-VIS2.0 DATASET

Method		Accuracy (%)	
CNN-L	Training Data	Original	69.11 ± 1.21
		LSSF	68.97 ± 1.24
	Network Fusion	Original+LSSF	79.96 ± 1.18
	Metric Learning	LDA (Original+LSSF)	$\mathbf{85.05 \pm 0.83}$
State-of-the-art	C-CBFD [29]		56.6 ± 2.4
	Dictionary Learning [23]		78.46 ± 1.67
	C-CBFD + LDA [29]		81.8 ± 2.3
	CNN + LDML [41]		85.9 ± 0.9
	Gabor + RBM [62]		$\mathbf{86.2 \pm 1.0}$

66.37 ± 1.45 to 68.97 ± 1.24, showing the value provided by our of bigger synthetic dataset.

4) Comparison With the State-of-the-Art: The CNN-L models in Table VIII are all trained using synthetic LFW data. First, LSSF-normalized and Original LFW synthetic data achieve very comparable performance: 68.97% vs. 69.11%. However, the fusion (averaging) of these two features can significantly improve the face recognition rates. It shows the fusion can keep the discriminative facial information but remove the illumination effects. Second, not surprisingly, metric learning can further improve the performance. The metric learning method used here is LDA, which is the most widely used one for face identification. Finally, Table VIII compares the proposed method against the state-of-the-art solutions [23], [29]. Reference [29] uses a designed descriptor that performs better in this dataset compared with other generic hand-crafted features, and LDA can further improve the accuracy. Our method significantly outperforms [29] when metric learning is not used (79.96% vs. 56.6%), and maintains superior performance when metric learning is used (85.05% vs. 81.8%). Reference [23] tries to solve the domain shift between two data sources (NIR and VIS) by a cross-modal metric learning: it assumes that a pair of NIR and VIS images shares the same sparse representation under two jointly learned dictionaries. Our method improves over [23] with a 7% margin without such an extra step of dictionary learning. Concurrently to our work, Saxena & Verbeek [41] obtained a comparable performance of 85.9% using a cross-modal version of LDML metric learning [11], albeit using CNN features learned from the 500 000 face images of the CASIA WEBFACE dataset. Finally, Yi *et al.* [62] obtained the best performance of 86.1% using an approach that extracts 40 dim. Gabor features at 176 local face regions, and uses these to train 176 different restricted Boltzmann machines (RBMs) specialized to model the modality shift at each face region. Note that unlike our approach based on feed-forward CNNs, their approach requires Gibbs sampling at test-time to infer the face representations.

VI. CONCLUSIONS AND FUTURE WORK

Recently, convolutional neural networks have attracted a lot of attention in the field of face recognition. However, deep learning methods heavily depend on big training data, which is not always available. To solve this problem in the field of face recognition, we propose a new face synthesis method which swaps the facial components of different face images to generate a new face. With this technique, we achieve state-of-the-art face recognition performance on LFW and CASIA NIR-VIS2.0 face databases.

In the future, we will apply this technique to more applications of face analysis. For example, the proposed data synthesis method can easily be used in training CNN-based face detection, facial attribute recognition, etc. More generally, the method applies to any objects which are well structured. For example, the human body is well structured and human images can be synthesised using this method. The synthetic images can be used to train deep models for pose estimation, pedestrian detection, and person re-identifiation.

ACKNOWLEDGMENT

The authors gratefully acknowledge NVIDIA for the donation of the GPUs for this research.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Prague, Czech Republic, 2004, pp. 469–481.
[2] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," in *Proc. Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–8.
[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[4] B.-C. Chen, C.-S. Chen, and W. Hsu, "Review and implementation of high-dimensional local binary patterns and its application to face recognition," Inst. Inf. Sci., Acad. Sinica, Taipei, Taiwan, Tech. Rep. TR-IIS-14-003, 2014.
[5] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. ECCV*, 2012, pp. 566–579.
[6] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. CVPR*, Jun. 2013, pp. 3025–3032.
[7] D. Decoste and B. Schölkopf, "Training invariant support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 161–190, 2002.
[8] C. Ding, J. Choi, D. Tao, and L. S. Davis. (2014). "Multi-directional multi-level dual-cross patterns for robust face recognition." [Online]. Available: https://arxiv.org/pdf/1401.5311.pdf
[9] Z.-H. Feng, G. Hu, J. Kittler, W. Christmas, and X.-J. Wu, "Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3425–3440, Nov. 2015.
[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, Jun. 2014, pp. 580–587.
[11] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. ICCV*, Sep. 2009, pp. 498–505.
[12] T. Hassner, S. Harel, E. Paz, and R. Enbar. (2014). "Effective face frontalization in unconstrained images." [Online]. Available: https://arxiv.org/abs/1411.7964
[13] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385
[14] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
[15] G. Hu *et al.*, "Attribute-enhanced face recognition with neural tensor fusion networks," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017.
[16] G. Hu *et al.*, "Face recognition using a unified 3D morphable model," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 73–89.
[17] G. Hu *et al.*, "Efficient 3D morphable face model fitting," *Pattern Recognit.*, vol. 67, pp. 366–379, Jul. 2017.
[18] G. Hu *et al.*, "When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 142–150.

[19] G. B. Huang *et al.*, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," College Inf. Comput. Sci., Univ. of Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007, vol. 1, no. 2.
[20] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. (2014). "Synthetic data and artificial neural networks for natural scene text recognition." [Online]. Available: https://arxiv.org/abs/1406.2227
[21] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. CVPR*, Jun. 2010, pp. 3304–3311.
[22] D. J. Jobson, Z.-U. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Trans. Image Process.*, vol. 6, no. 3, pp. 451–462, Mar. 1997.
[23] F. Juefei-Xu, D. Pal, and M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 141–150.
[24] S. G. Kong, J. Heo, B. R. Abidi, J. Paik, and M. A. Abidi, "Recent advances in visual and infrared face recognition—A review," *Comput. Vis. Image Understand.*, vol. 97, no. 1, p. 103–135, 2005.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[26] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2013, pp. 348–353.
[27] J. Liu, Y. Deng, and C. Huang. (2015). "Targeting ultimate accuracy: Face recognition via deep embedding." [Online]. Available: https://arxiv.org/abs/1506.07310
[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[29] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 2041–2056, Oct. 2015.
[30] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *Proc. ECCV*, 2016, pp. 579–596.
[31] D. Miller, E. Brossard, S. M. Seitz, and I. Kemelmacher-Shlizerman. (2015). "Megaface: A million faces for recognition at scale." [Online]. Available: https://arxiv.org/abs/1505.02108
[32] S. Ouyang, T. Hospedales, Y.-Z. Song, and X. Li. (2016). "A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution." [Online]. Available: https://arxiv.org/abs/1409.5114
[33] J. Papon and M. Schoeler. (2015). "Semantic pose using deep networks trained on synthetic RGB-D." [Online]. Available: https://arxiv.org/abs/1508.00835
[34] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, p. 6.
[35] M. Paulin, J. Revaud, Z. Harchaoui, F. Perronnin, and C. Schmid, "Transformation pursuit for image classification," in *Proc. CVPR*, Jun. 2014, pp. 3646–3653.
[36] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, 2003.
[37] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. CVPR*, Jun. 2015, pp. 1713–1721.
[38] G. Rogez and C. Schmid, "MoCap-guided data augmentation for 3D pose estimation in the wild," in *Proc. NIPS*, 2016, pp. 3108–3116.
[39] A. Rozantsev, V. Lepetit, and P. Fua, "On rendering synthetic images for training an object detector," *Comput. Vis. Image Understand.*, vol. 137, pp. 24–37, Aug. 2015.
[40] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.
[41] S. Saxena and J. Verbeek, "Heterogeneous face recognition with CNNs," in *Proc. ECCV TASK-CV Workshop*, 2016, pp. 483–491.
[42] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, Jun. 2015, pp. 815–823.
[43] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," *Commun. ACM*, vol. 56, no. 1, pp. 116–124, 2013.
[44] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2002, pp. 53–58.
[45] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild." in *Proc. Brit. Mach. Vis. Conf.*, 2013, p. 4.
[46] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. NIPS*, 2014, pp. 568–576.
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.
[48] H. Su, C. Qi, Y. Li, and L. J. Guibas, "Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views," in *Proc. ICCV*, Dec. 2015, pp. 2686–2694.
[49] C. Sun and R. Nevatia, "ACTIVE: Activity concept transitions in video event classification," in *Proc. ICCV*, Dec. 2013, pp. 913–920.
[50] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.
[51] Y. Sun, D. Liang, X. Wang, and X. Tang. (2015). "DeepID3: Face recognition with very deep neural networks." [Online]. Available: https://arxiv.org/abs/1502.00873
[52] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *Proc. ICCV*, Dec. 2013, pp. 1489–1496.
[53] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1891–1898.
[54] Y. Sun, X. Wang, and X. Tang. (2014). "Deeply learned face representations are sparse, selective, and robust." [Online]. Available: https://arxiv.org/abs/1412.1265
[55] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Jun. 2015, pp. 1–9.
[56] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. CVPR*, Jun. 2014, pp. 1701–1708.
[57] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 1991, pp. 586–591.
[58] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 113, pp. 371–408, Dec. 2010.
[59] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.
[60] M. Weinmann, J. Gall, and R. Klein, "Material classification based on training data synthesized using a BTF database," in *Proc. ECCV*, 2014, pp. 156–171.
[61] X. Xie, W.-S. Zheng, J. Lai, P. C. Yuen, and C. Y. Suen, "Normalization of face illumination based on large- and small-scale features," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1807–1821, Jul. 2011.
[62] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, May 2015, pp. 1–7.
[63] D. Yi, Z. Lei, S. Liao, and S. Z. Li. (2014). "Learning face representation from scratch." [Online]. Available: https://arxiv.org/abs/1411.7923
[64] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Sketch-a-Net: A deep neural network that beats humans," *Int. J. Comput. Vis.*, vol. 122, no. 6, pp. 411–425, 2017.
[65] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.
[66] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Proc. ECCV*, 2014, pp. 94–108.
[67] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. (2015). "Face alignment across large poses: A 3D solution." [Online]. Available: https://arxiv.org/abs/1511.07212
[68] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 787–796.

Guosheng Hu, photograph and biography not available at the time of publication.

Xiaojiang Peng, photograph and biography not available at the time of publication.

Yongxin Yang, photograph and biography not available at the time of publication.

Timothy M. Hospedales, photograph and biography not available at the time of publication.

Jakob Verbeek, photograph and biography not available at the time of publication.