DoppelVer: A Benchmark for Face Verification

Nathan Thom, Andrew DeBolt, Lyssie Brown, and Emily M. Hand

University of Nevada, Reno, Reno NV 89557, USA
{nthom,adebolt,abrown}@nevada.unr.edu
emhand@unr.edu
https://www.machineperceptionlab.com

Abstract. The field of automated face verification has become saturated in recent years, with state-of-the-art methods outperforming humans on all benchmarks. Many researchers would say that face verification is close to being a solved problem. We argue that evaluation datasets are not challenging enough, and that there is still significant room for improvement in automated face verification techniques. This paper introduces the DoppelVer dataset, a challenging face verification dataset consisting of doppelganger pairs. Doppelgangers are pairs of individuals that are extremely visually similar, oftentimes mistaken for one another. With this dataset, we introduce two challenging protocols: doppelganger and Visual Similarity from Embeddings (ViSE). The doppelganger protocol utilizes doppelganger pairs as negative verification samples. The ViSE protocol selects negative pairs by isolating image samples that are very close together in a particular embedding space. In order to demonstrate the challenge that the DoppelVer dataset poses, we evaluate a state-of-the-art face verification method on the dataset. Our experiments demonstrate that the DoppelVer dataset is significantly more challenging than its predecessors, indicating that there is still room for improvement in face verification technology.

Keywords: face verification · datasets · negative pair selection.

1 Introduction

The task of face recognition has received considerable attention from computer vision and pattern recognition researchers in the past 20 years. This is because face identification has significant utility in the fields of biometrics, visual search, and socially assistive technologies [1,11]. Additionally, compute equipment capable of running increasingly powerful algorithms has become relatively cheap and widely available. Face recognition technologies have significant impact on society with a market share of $5.69 billion worldwide in 2023 and a projected $12.05 billion by 2028 [18].

Work in face recognition and verification is dataset motivated. Every time a new dataset is released, there are significant improvements in face verification technology. Over the last several decades, there have been many datasets which have challenged the state-of-the-art (SOTA) face verification methods, such as

Labeled Faces in the Wild (LFW), IARPA Janus Benchmarks A, B, and C (IJB-{A,B,C}), etc. [8,10,24,17]. With the release of these datasets came a renewed interest in the field. Over the last few years, however, face identification on these datasets has reached a saturation point. For example, many methods achieve over 99% accuracy on the LFW benchmark. With such high accuracies we are able to visually inspect the samples that are incorrectly classified. In many cases these incorrectly classified samples are mislabeled meaning there is really no room for improvement on these datasets. In addition, face identification datasets are often collected with a focus on quantity, neglecting other important attributes. These problems provide the motivation for the proposed work.

This report introduces a new dataset – *DoppelVer* – consisting of unconstrained face images of doppelgangers – that is, individuals who look very similar and are often mistaken for each other. The purpose of DoppelVer is to challenge current SOTA facial feature extraction and face verification and identification methods. Although a plethora of datasets have been published to this end in the past decade, many of them are either unavailable or have been nearly solved. DoppelVer offers a specific challenge for modern face recognition methods, specifically the task of differentiating individuals who could pass for each other. To the best of our knowledge DoppelVer is the first dataset to increase face classification difficulty by increasing inter-class similarity rather than decreasing intra-class similarity. Upon publication of this paper, DoppelVer will be made publicly available.

Here we detail the highlights of the DoppelVer dataset, which will be expanded upon in the remainder of this work.

- DoppelVer contains 390 unique identities, each with at least one corresponding doppelganger pair.
- We provide the unaltered source images along with cropped, aligned, and centered (CCA) images.
- There is an average of 72 CCA samples per identity, with a minimum of 11 and a maximum of 98.
- For the CCA images we provide two evaluation protocols: doppelganger and **Vi**sual **S**imilarity from **E**mbeddings (ViSE). Under the doppelganger protocol negative samples are select images depicting an identity's doppelganger. The ViSE protocol uses a generalized image embedding model to select negative images that are highly visually similar to the current image sample.
- Both protocols are divided into 10 cross validation splits which are distinct across identities. The doppelganger protocol's cross validation splits are made up of 14,000 image pairs while ViSE's splits contain 3,500 image samples.

The remainder of the paper is organized as follows: in Section 2 we provide background to the field of face recognition, with a focus on feature extraction and face classification methods. Section 2 also details similar datasets and the novelty of DoppelVer. Section 3 contains a more detailed description of the DoppelVer dataset including data collection, pre-processing, labeling, and the generation of the evaluation protocols. In Section 4 we provide results of our experimentation

comparing the performance of SOTA facial recognition pipelines on existing benchmark datasets and DoppelVer.

2 Related Work

2.1 Background

Face recognition is separated into three well-defined steps: (1) face detection and localization, (2) extraction of features from the detected face, and (3) classification (verification or identification) [11]. The first task is to decide whether or not there are faces in an image. If there are one or more faces, then the system identifies bounding boxes for each face. The feature extraction step generates a feature vector from the localized face. This feature vector should be discriminative enough to separate images of one identity from images of other identities. Lastly, there is the classification step. This is separated into two classes of techniques: identification and verification. In the identification scenario the system is aware of a finite number of identities and it should learn to match each image sample to one identity class. For the verification task the model is only provided with supervision in the form of a binary label which represents either same or different, and so pairs of images are compared at each step.

Any face recognition system that is meant to be deployed in "the wild" will need to perform all three of these steps. That being said, each step is commonly considered an active research topic. The intended purpose of the DoppelVer dataset is to contribute towards improvements in the final two steps. In this work, we devote our efforts towards the feature extraction and classification tasks. This is because most modern methods employ deep learning techniques, which combine feature extraction and classification into a single system. Additionally, research has seemingly slowed in these areas.

One might suggest that the field of face classification is reaching its maturity, citing results on the well-known benchmarks such as LFW, AgeDB, or IJB-{A,B,C} [8,19,10,24,17]. Rather than assuming that the reported metrics are due to the techniques solving the task of visually recognizing faces, we hypothesize that the modern techniques have improved beyond the level of difficulty provided by the current benchmarks. For example, in 2015 Liu et al. published a result of 99.77% accuracy on the LFW benchmark [13]. The dataset's evaluation protocol contains only 6000 images. This means that for nearly a decade methods have been attempting to show improvements on a method that mis-classifies only 14 images, five of which are known to be incorrectly labeled.

Other methods have emerged with the intent of contributing to the issue of increasing unconstrained face recognition benchmark difficulty [21,19,29,28]. These methods primarily focus on increasing difficulty of the classification task with highly varied pose and age. These features essentially decrease the intra-class similarity (i.e. selecting images of the same identity that are visually different).

Our DoppelVer dataset increases classification difficulty by increasing inter-class similarity (i.e. selecting images of different identities that are visually sim-

ilar). We accomplish this goal in two distinct ways. First, we aggregate doppelganger pairs. A doppelganger pair is simply two individuals who have similar facial features. This protocol is constructed by human labelers selecting visually similar identities. Second, for a given image we mine a negative sample which is highly visually similar. This is accomplished by generating an embedding or latent vector for all images in the dataset. We search for pairs of images whose embeddings are near one another in the latent space. By these two methods we produce two protocols that we have named doppelganger and **Vi**sual **S**imilarity from **E**mbeddings (ViSE).

2.2 Existing Datasets

There are a large number of datasets collected and presented for the purpose of facial feature extraction and classification. Many of these datasets are designed either for training or evaluation. Here we describe the major datasets that already exist for the purpose of model evaluation and benchmarking and compare them with the proposed DoppelVer dataset.

Labeled Faces in the Wild (LFW) [8]: The LFW dataset was created by Huang et al. in 2007. At the time of publishing, many face recognition datasets were collected by small teams of researchers with the intent of collecting facial images in constrained settings. LFW however was meant for studying the problem of recognizing faces in unconstrained settings. The dataset contains 13,233 images and 7,549 identities. The researchers behind LFW contributed significantly to the field by presenting a dataset organization that focused on the honest reporting of results for the task of open-set face recognition. Their dataset contains a development view and an evaluation view as well as splits for 10 fold cross-validation. The current SOTA accuracy on LFW is 99.8% ($\pm$0.2001) [2].

AgeDB [19]: This dataset was introduced in 2017, with a focus on accurate hand-labeling of age. This is a useful database when performing tasks such as age-invariant face verification, age estimation, and face age progression. The database contains 16,488 images of 568 identities with accurate-to-the-year age labels. The average number of images per individual is 29, with an age range of 1 to 101 years old, the average age for an individual being 50.3 years. AgeDB provides four face verification protocols, each split into 10 folds following LFW's process. These four protocols restrict the age variance across sample pairs. The provided protocols cap age range to 5, 10, 20 and 30 years respectively. The current SOTA accuracy on AgeDB 30 is 98.7% [3]

Cross-Age LFW (CA-LFW) [29]: The authors of this database posit that methods reporting accuracy on LFW's benchmark are optimistic. To show this, CA-LFW has both positive and negative pairs which depict a large age gap, while also providing negative pairs which are of the same race and gender. These visually similar negative pairs emphasize the effect of age difference on classifier performance. This dataset contains the same identifies as LFW with 6,000 image pairs. The current SOTA accuracy on CA-LFW is 95.87% [5]

Cross-Pose LFW (CP-LFW) [28]: CP-LFW was proposed by the same authors as CA-LFW and was released one year later. This publication shifts focus to the important task of face verification in the presence of extreme pose. They note that nearly all images in LFW are near-frontal, suggesting that results on LFW provide a poor representation of a face recognition method's performance when deployed into a real setting. The current SOTA accuracy on CA-LFW is 92.08% [5]

Each of the databases detailed above provide an important contribution to furthering the field of face recognition. These datasets provide unconstrained images and in the cases of [21,19,29,28] the sample pairs vary along specific axis which were not well represented in LFW. As mentioned previously, these datasets focus on selecting positive pairs which are visually dissimilar to one another. DoppelVer's goal is to expand on a dimension of challenge which has not yet been addressed. This dimension is that of visual similarity among negative samples. This yet unseen challenge will force methods to extract significantly more fine-grained, prominent features from face images. In order to achieve high performance on DoppelVer, techniques will be required to extract those features which uniquely define a given identity.

3 Proposed Method

3.1 Dataset Collection

In order to construct a dataset for which negative samples are analogous to positive samples it is intuitive to begin by aggregating a list of identities which bare visual similarity to human labelers (i.e. doppelgangers). Doppelganger identity pairs were collected through labeler intuition of similar looking identities and lists of doppelgangers publicly available on the Internet. We present a large list of doppelganger identity combinations, totalling 237 pairs and 390 individuals. For each individual, 100 images were scraped from online sources. The average number of images presented in the dataset for each person is approximately 72 due to pruning of noisy samples and duplicates.

3.2 Data Preparation

Data preparation involved two distinct steps: (1) cropping, aligning and centering the images, and (2) hand removal of erroneous samples and duplicate images.

Cropping, aligning and centering: The first step in the data preparation is to reduce the original images into cropped, aligned, and centered images. We crop to remove information which is extraneous to the face recognition task. Alignment and centering are performed as they have been recognized as important for achieving competitive face recognition benchmark performance. Alignment involves rotating the image such that the eyes lie on a horizontal line (i.e. the same y-coordinates). The operation of centering moves the face in the frame of

the image such that it appears centrally. Centering is accomplished by repeating edge pixels along either the horizontal or vertical borders of the image. The cropping operation relies on a bounding box and centering/alignment rely on facial landmarks. We extract the bounding boxes and facial landmarks for images in DoppelVer with the MTCNN detector [27].

While processing the dataset with MTCNN, three cases may occur: (1) MTCNN does not detect a face, (2) MTCNN detects a single face, and (3) MTCNN detects multiple faces. Images where a face is not detected are pruned from the dataset. Although MTCNN returns a detected face in most images, not all detections contain the target identity or a valid face. Each detection is hand-checked for validity during the cleaning phase of pre-processing. When at least one face is detected, MTCNN returns a bounding box for the image along with five facial landmarks. The landmarks provide the detected location of the centers of the eyes, corners of the mouth, and tip of the nose.

Initially we cropped the source images to the bounding boxes predicted by MTCNN, but found that the crop was too tight. These crops often removed valuable information such as the top of head, ears and most of the neck. We expand MTCNN's detected bounding box width and height by 50%. This produces crops which contain more contextual information. There are cases for which the detected face is near the border of the image, restricting our ability to expand the bounding box. In these cases we simply set the desired bounding box location to the border of the image.

After cropping, we align the images according to the extracted landmark locations. Our alignment rotates the images such that the detected landmark for left and right eyes have the same y-axis coordinate. During the alignment process some image information is lost due to the corners of the image rotating outside of the frame. Following the lead of the CelebA dataset, we reduce the effects of this information loss by performing same padding for any pixels that are lost due to rotation [16].

The last pre-processing step is to center the image so that the center most pixel of the image is within the bounds of the detected face. Centering is performed by computing a landmark which lies at the mid-point between the left and right eye landmarks. Additional pixels are appended to the horizontal and vertical image borders such that the center of the face is equidistant to each border. The appended pixels are simply duplicates of the pixels which are along the border that needs to be expanded.

Removal of erroneous or duplicate Images: We remove unsatisfactory images by hand and by automatic detection. In the case of hand labeling, labelers began with the original image set collected from the internet. Their task was to pass over the images and delete any image which contained erroneous detections (e.g. not depicting the correct identity or images not containing a face). The set of images which had complete labeler agreement was accepted. The set of images which did not have agreement were re-labeled. Any remaining images which the labelers did not reach agreement on were pruned from the dataset. The images

which achieved hand label agreement were passed to the automatic detection system.

The automatic detection system works by generating embeddings for each face image in the dataset with the dinov2s model [20]. dinov2s is a general purpose image embedding model, built to capture a discriminative representation of input images without finetuning. The cosine similarity is computed between all combinations of input images' embeddings to determine samples which are highly visually similar. To compute the embeddings and cosine similarities efficiently we utilize the fastdup library [12] from Visual Layer. For any image pair that has exact similarity (i.e. duplicate images), one image from the pair is pruned from the dataset. Next, we return all of the image pairs that are above a threshold of 0.92 similarity. We extract these images pairs and provide them to human labelers to find near duplicate images (i.e. images that have been horizontally flipped, color jittered, cropped slightly differently, etc.), which are removed from the dataset.

3.3 Protocol Generation

The DoppelVer dataset contains in total 27,967 carefully curated and processed images. The question that remains is the best way to utilize these images for assessing and benchmarking feature extraction and face classification methods. To answer this question, we introduce two protocols for evaluation using DoppelVer: doppelganger and ViSE. Fig. 1 provides example image pairs for each protocol in DoppelVer and Fig. 2 shows samples from CA-LFW and CP-LFW.

Both protocols are made up of positive and negative image pairs. Positive image pairs in both protocols signify instances where both images depict the same identity. In the doppelganger protocol, negative pairs are made up of one image sample depicting the current target identity and one image sample depicting their doppelganger. In the ViSE protocol the negative pairs contain an image sample depicting an identity which does not generally appear as visually similar to the current identity, but in a one-off case is visually similar. Such similarity often arises due to comparable pose, lighting, hair style, clothing, or image background. After generating a large number of image pairs, we divide the dataset into 10 equally sized splits. Each split is divided such that images of an identity are in only a single split. Identities are divided the same in each protocol (e.g. split 0 of the doppelganger protocol depicts the same identities as split 0 of ViSE).

The doppelganger protocol is generated with our curated list of doppelganger pairs. We create the pair instances in the doppelganger protocol as follows. First, we sample 500 image combinations, without replacement, for every pair of doppelgangers and identities with themselves. After generating all pairs following this criteria we separate the samples into 10 splits based on their identities and pairs such that the same identity never shows up in multiple splits. Approximately 10 percent of the dataset is placed into each split. Finally, from each split we randomly sample 7,000 positive pairs and 7,000 negative pairs. We do this to follow the procedures laid out by LFW. This protocol has a positive label

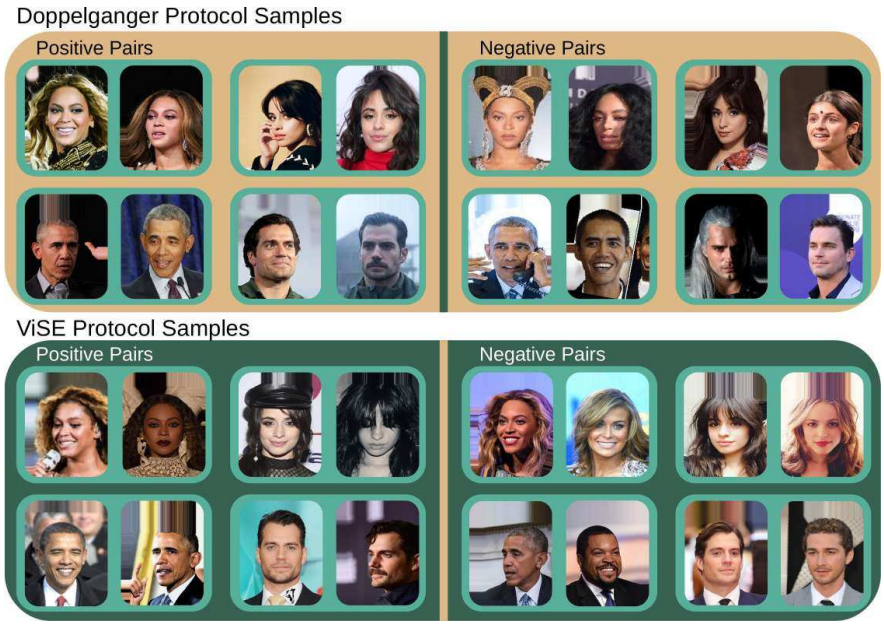

Fig. 1. Shown above are samples from both protocols of the DoppelVer dataset – doppelganger and ViSE. We note that negative samples from the Doppelganger protocol share facial attributes while the image pairs in ViSE frequently share factors external to the face such as pose, clothing, and background.

Fig. 2. The upper portion of this figure presents samples from the CA-LFW dataset and the lower portion contains samples from CP-LFW. The CA-LFW samples showcase differences in age while CP-LFW's images showcase differences in pose.

and negative label ratio of exactly 50%. It has a gender distribution of 44.96% males and 55.04% female samples respectively. Identities in each split have a relatively even representation with an average minimum contribution of 4.31%, average maximum contribution of 19.07%, and an average standard deviation between representation of 5.32%. In total the doppelganger protocol has 140,000 sample pairs.

To generate the ViSE protocol we use a similar approach to the one described in the automatic detection of unsatisfactory images. We begin by generating embeddings for each image in the dataset with the dinov2s model. Next, we compute the cosine similarity between images which do not come from the same identity. We retain all image pairs that have a similarity greater than 0.80. We have found that this form of mining hard pairs image by image rather than individual by individual results in significantly more visual similarity between image pairs. Using the same identities in each split as the doppelganger protocol, we break the protocol into 10 splits with unique identities in each split. This protocol has a positive label and negative label ratio of exactly 50%. It has a gender distribution of 40.36% male and 59.64% female. Identities in each split have a relatively even representation with an average minimum contribution of 2.29%, average maximum contribution of 17.61%, and an average standard deviation between representation of 3.6%. This protocol has 35,000 verification pairs.

3.4 Intended Use

The DoppelVer dataset is intended to provide a new challenge for the research community developing methods in the area of facial recognition. DoppelVer has been designed to act as an evaluation dataset, not a training dataset. In the past decade the most effective methods of facial recognition have utilized large training sets such as CASIA-WebFace, MegaFace, VGGFace2, MS-Celeb-1M [26,9,4,7]. These datasets contain 34.94K, 1.03M, 3.31M, 10M samples respectively. Although an aggregate of visually difficult pairs is attractive for faster convergence time, DoppelVer does not contain enough diversity to effectively and ethically train models.

We provide cross validation splits for both protocols in DoppelVer. The purpose of these splits is two-fold. First, some methods may wish to perform feature extraction prior to face classification. Such extraction methods should pre-train on external sources and infer features for each image in DoppelVer. At evaluation time final-stage classifiers should be iteratively trained from scratch (using their pre-trained feature extraction methods) on nine splits and evaluated on the tenth. Performance should be recorded as an average across the 9 models. We refer to interaction with the dataset in this way as **View 1**. Second, methods that wish to train on external data and perform only evaluation on DoppelVer should use split 0 for algorithm development and validation of results. The model should not be exposed to data in any of the other nine splits until final evaluation. Use of the dataset in this way is called **View 2**.

Taking motivation from the LFW dataset, we suggest that researchers utilizing **View 1** report estimated mean accuracy (EM ACC) and standard error of the mean (SEM). We define these metrics in the following way:

$$\hat{\mu} = \frac{\Sigma_{i=1}^{9} p_i}{9}, SEM = \frac{\hat{\sigma}}{\sqrt{9}}, \hat{\sigma} = \sqrt{\frac{\Sigma_{i=1}^{9}(p_i - \hat{\mu})^2}{9}}$$

where p_i is the percentage of correct classifications on **View 1** when using the i^{th} split for testing. $\hat{\sigma}$ is the estimate of the standard deviation. As noted by the authors of LFW, it is important that accuracy is computed with parameters and thresholds chosen independently of the test data. Researchers should not simply choose the point on a Precision-Recall curve giving the highest accuracy.

For the methods which utilize **View 2** of DoppelVer, we advocate for the use of accuracy (ACC) and area under the receiver operating characteristic curve (ROC AUC). We elect for the use of ACC and ROC AUC because of the balanced nature of classes in the Doppelganger and ViSE protocols. In addition, the correct classification of true positives is equally important to classification of true negatives.

4 Experiments

In this section, we highlight the challenges posed by the DoppelVer dataset as compared to other existing evaluation datasets. We detail the methods used for evaluation, the training data, and the process employed for training and testing.

4.1 Evaluation Model

To provide an accurate depiction of the challenge posed by DoppelVer, it is important that we evaluate DoppelVer with SOTA face recognition models. Due to ease of implementation and competitive results we have elected to utilize the techniques described by Wen et al. in SphereFace2 [23]. In particular we train the 20 layer SphereFace Network (SFNet-20), initially proposed in [14], with the following loss functions: COCO, SphereFace, CosFace, ArcFace, and SphereFace2. Following Wen et al., we equip SFNet-20 with batch normalization to facilitate model optimization. A complete implementation for training SFNet-20 with the aforementioned loss functions can be found in the OpenSphere GitHub repository [25].

4.2 Training and Evaluation Process

For pre-processing, we crop face images in each dataset with MTCNN, resize images to a size of 112×112, and normalize each RGB pixel [0, 255] to the range [-1, 1]. We trained our models on a single Nvidia Geforce RTX 3090 GPU. Each model is trained for 70,000 batches of size 512. The model weights are updated by stochastic gradient descent with a momentum of 0.9 and weight decay of

0.0005. The initial learning rate of 0.1 is reduced by a factor of 0.1 at batches 40,000; 60,000; and 70,000.

We evaluate our dataset and protocols with VGGFace2, MS-Celeb-1M, and CASIA-WebFace [4,7,26]. In each run the VGGFace2 dataset was found to produce the best results on each evaluation dataset. VGGFace2 contains between 80 and 800 images for each identity making it a powerful training dataset for the face verification task. Evaluation of the trained models is performed on LFW, CA-LFW, CP-LFW, AgeDB 30, view 2 of DoppelVer's doppelganger protocol, and view 2 of DoppelVer's ViSE protocol. Our measured accuracy and ROC AUC are provided in Tables 1 and 2 respectively.

4.3 Discussion of Results

We are satisfied with the performance achieved by the SOTA methods on the existing benchmark datasets. SOTA performance on the LFW dataset is 99.8% accuracy. Our training of SphereFace achieves an accuracy of 99.58%, misclassifying just 25 samples. With this result we can be assured that this baseline is competitive with other SOTA methods. The best published results on the other benchmark datasets are 95.87%, 92.08%, and 98.7% accuracy on CA-LFW, CP-LFW, and AgeDB 30 respectively. Regardless of loss function, the baseline networks struggle significantly more with variations in pose than variations in age. CA-LFW and AgeDB appear to present a similar degree of difficulty to the models.

It is clear from our experiments that the doppelganger and ViSE protocols of DoppelVer are much more difficult for the classifiers than the other datasets.

Table 1. Average accuracy of face verification for the comparison models trained with VGGFace2 and benchmarked on various datasets.

Method	LFW	CA-LFW	CP-LFW	AgeDB	Doppelganger	ViSE
COCO [15]	99.08	91.25	88.48	89.40	61.14	52.53
SphereFace [14]	99.58	93.15	91.65	93.53	63.48	57.08
CosFace [22]	99.52	93.03	91.37	93.02	63.29	56.93
ArcFace [6]	99.55	93.40	91.18	92.57	63.28	57.70
SphereFace2 [23]	99.53	93.80	90.83	93.38	61.66	55.41
Average	**99.45**	**92.93**	**90.70**	**92.38**	**62.57**	**55.93**

Table 2. Average AUC of face verification for the comparison models trained with VGGFace2 and benchmarked on various datasets.

Method	LFW	CA-LFW	CP-LFW	AgeDB	Doppelganger	ViSE
COCO [15]	99.89	96.56	93.57	96.03	65.13	50.53
SphereFace [14]	99.92	97.44	95.50	98.11	68.65	59.41
CosFace [22]	99.91	97.28	95.64	97.86	67.91	58.58
ArcFace [6]	99.89	96.99	95.46	97.53	68.15	59.79
SphereFace2 [23]	99.89	97.55	95.42	98.02	65.43	55.77
Average	**99.90**	**97.16**	**95.12**	**97.51**	**67.05**	**56.82**

Results are better for the doppelganger protocol than the ViSE protocol. This result aligns with intuition. Two identities that are doppelgangers may in general share facial attributes, but variations in clothing, hair style, lighting, and facial expression are expected when viewing a gallery of images depicting them.

On the other hand, the ViSE protocol contains image pairs which are adversarial in nature. By this we mean that the combinations of samples are those which a deep network is expected to struggle to differentiate. Although we use a different deep convolutional network to select samples which are visually similar than we do for performing facial recognition, one would expect that the visual features which are attended to by deep networks would have some similarity.

We believe that methods which will perform well on the ViSE protocol will need to extract features which are highly specific to the task of facial recognition. In addition, methods will need to not only detect relevant facial features, but discern if the features are prominent/defining to the individual's face.

5 Conclusion

In this work we introduce DoppelVer, a novel evaluation dataset for the tasks of facial feature extraction and face verification. DoppelVer consists of 27,967 carefully curated face images, which are used in two face verification protocols of image pairs: doppelganger and ViSE. We evaluate our methods using several SOTA methods. A near SOTA baseline model is only capable of correctly performing face verification at an accuracy of 62.57% and 55.93% in the doppelganger and ViSE protocols respectively. This indicates that despite impressive results on popular benchmark datasets, there is still work to be done in the field of facial recognition.

Future research should explore improvements to deep vision models to enable accurate classification of visually similar individuals. Additionally, future work might involve the application of the ViSE protocol's adversarial image pair selection to larger selections of facial data to enable the training of deep networks with visually similar negative pairs. Lastly, this data might be used to understand the difference in vision model perceptions between images of identical twins or parents and children at similar times of life.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grants No. 1909707, 2150394, and 2302187. Standard disclaimers apply.

References

1. Adjabi, I., Ouahabi, A., Benzaoui, A., Taleb-Ahmed, A.: Past, present, and future of face recognition: A review. Electronics **9**(8) (2020). https://doi.org/10.3390/electronics9081188, https://www.mdpi.com/2079-9292/9/8/1188
2. Alansari, M., Hay, O.A., Javed, S., Shoufan, A., Zweiri, Y., Werghi, N.: Ghostfacenets: Lightweight face recognition model from cheap operations. IEEE Access **11**, 35429–35446 (2023). https://doi.org/10.1109/ACCESS.2023.3266068
3. An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., Wu, L., Qin, B., Zhang, M., Zhang, D., et al.: Partial fc: Training 10 million identities on a single machine. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1445–1449 (2021)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018). pp. 67–74 (2018). https://doi.org/10.1109/FG.2018.00020
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
6. Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 5962–5979 (2022). https://doi.org/10.1109/TPAMI.2021.3087709
7. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 87–102. Springer International Publishing, Cham (2016)
8. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
9. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4873–4882 (2016). https://doi.org/10.1109/CVPR.2016.527
10. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1931–1939 (2015). https://doi.org/10.1109/CVPR.2015.7298803
11. Kortli, Y., Jridi, M., Al Falou, A., Atri, M.: Face recognition systems: A survey. Sensors **20**(2) (2020). https://doi.org/10.3390/s20020342, https://www.mdpi.com/1424-8220/20/2/342
12. visual layer: fastdup. https://github.com/visual-layer/fastdup (July 2023)
13. Liu, J., Deng, Y., Bai, T., Wei, Z., Huang, C.: Targeting Ultimate Accuracy: Face Recognition via Deep Embedding. arXiv e-prints arXiv:1506.07310 (Jun 2015). https://doi.org/10.48550/arXiv.1506.07310
14. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6738–6746 (2017). https://doi.org/10.1109/CVPR.2017.713

15. Liu, Y., Li, H., Wang, X.: Rethinking feature discrimination and polymerization for large-scale recognition. CoRR **abs/1710.00870** (2017), http://arxiv.org/abs/1710.00870
16. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 3730–3738 (2015). https://doi.org/10.1109/ICCV.2015.425
17. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., Grother, P.: Iarpa janus benchmark - c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB). pp. 158–165 (2018). https://doi.org/10.1109/ICB2018.2018.00033
18. MordorIntelligence: (July 2023), https://www.mordorintelligence.com/industry-reports/facial-recognition-market
19. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop. vol. 2, p. 5 (2017)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. arXiv e-prints arXiv:2304.07193 (Apr 2023). https://doi.org/10.48550/arXiv.2304.07193
21. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9 (2016). https://doi.org/10.1109/WACV.2016.7477558
22. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018). https://doi.org/10.1109/CVPR.2018.00552
23. Wen, Y., Liu, W., Weller, A., Raj, B., Singh, R.: Sphereface2: Binary classification is all you need for deep face recognition. CoRR **abs/2108.01513** (2021), https://arxiv.org/abs/2108.01513
24. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., Cheney, J., Grother, P.: Iarpa janus benchmark-b face dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 592–600 (2017). https://doi.org/10.1109/CVPRW.2017.87
25. ydwen: Opensphere. https://github.com/ydwen/opensphere (July 2023)
26. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. CoRR **abs/1411.7923** (2014), http://arxiv.org/abs/1411.7923
27. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016). https://doi.org/10.1109/LSP.2016.2603342
28. Zheng, T., Deng, W.: Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Tech. Rep. 18-01, Beijing University of Posts and Telecommunications (February 2018)
29. Zheng, T., Deng, W., Hu, J.: Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. CoRR **abs/1708.08197** (2017), http://arxiv.org/abs/1708.08197