Harnessing visual context information to improve face identification accuracy and explainability

by

Hai Phan

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 4, 2024

Keywords: transformer, vision transformer, earth mover distance, structure similarity, cosine similarity

Approved by

Anh Nguyen, Chair, Assistant Professor of Computer Science and Software Engineering
Pan He, Assistant professor of Computer Science and Software Engineering,
Yang Zhou, Assistant Professor of Computer Science and Software Engineering
Santu Karmaker, Assistant Professor of Computer Science and Software Engineering

Abstract

Face identification (FI) is ubiquitous and drives many high-stake decisions made by the law enforcement. A common FI approach compares two images by taking the cosine similarity between their image embeddings. Yet, such approach suffers from poor out-of-distribution (OOD) generalization to new types of images (e.g., when a query face is masked, cropped or rotated) not included in the training set or the gallery. Recently, interpreatable deep metric learning with structural matching (e.g. DIML [101] and Vision Transformers [27]) obtained significant outcomes in popular computer vision problems such as image classification, image clustering, etc. In this proposal, we present simple yet efficient schemes to exploit structural similarity for an interpretable face matching algorithms. We propose two following novel methods.

- DeepFace-EMD [63]: A re-ranking approach that compares two faces using the Earth Mover's Distance on the deep, spatial features of image patches.
- Face-ViT [62]: A novel architectural design using Vision Transformers (ViTs) for out-of-distribution (OOD) face identification and show significant improvement in inference speed. We feed embeddings of both images through a pre-trained CNN by ArcFace [22], layers of a Transformer encoder, and two linear layers as part of a ViT. We train the model with 2M pairs sampled from the CASIA Webface [93]

Our extra comparison stage explicitly examines image similarity at a fine-grained level (e.g., eyes to eyes) and is more robust to OOD perturbations and occlusions than traditional FI. Interestingly, without finetuning feature extractors, our method consistently improves the accuracy on all tested OOD queries: masked, cropped, rotated, and adversarial while obtaining similar results on in-distribution images. Moreover, our model demonstrates significant interoperability through the visualization of cross-attention.

Acknowledgements

As my PhD comes to a close, I want to thank all of those who have made my research possible and my time enjoyable. First and foremost is my advisor, Prof. Anh Nguyen, for many years of advice and guidance throughout my academic career. From my time in the lab, as a Ph.D. student, and as a teaching assistant, I have gained immeasurable experience that will no doubt help guide me through the rest of my career path. I am immensely grateful to Prof. Xiao Qin for his generous support and invaluable guidance throughout the completion of my PhD. Additionally, his efficient assistance in successfully navigating the qualifying exams was instrumental in my progress. I also wish to thank my doctoral committee members, Prof. Pan He, Prof. Yang Zhou, and Prof. Santu Karmaker as well as university reader Prof. Shiwen Mao for taking the time to be a part of the completion of my Ph.D.

All the members of the lab center throughout the years that I have been there deserve thanks for all the help they have provided along the way: Thang Pham, Giang Nguyen, Qi Li, Peijie Chen, Mohammad Taesir for valuable suggestions. I am sure all of you will go on to do even more impressive research than you have already done and I wish you all the best.

I would like to express my heartfelt gratitude to my co-authors Cindy Le, Vu Le, and my former colleague at Carnegie Mellon University, Ethan (Yihui) He, for their invaluable contributions in polishing and improving my latest research work. I have great confidence in your future successes in your respective careers.

I would like to express my gratitude to my former mentors: Dr. Kai Li (Meta) and Dr. Hao Tan & Dr. Trung Bui (Adobe) for the exceptional internships. I thoroughly enjoyed my time with you and gained invaluable experience, acquiring essential skills for working in the industry.

I also must thank Auburn staff who have always helped us with any problems or last-minute requests (of which there were quite a few). So thank you, Kelly, Jo Ann, and anyone else who I may have forgotten.

My family has always been a source of strength and comfort for me over the years and I feel exceptionally lucky that my mother, sister, and aunt remained always within reach, sometimes literally, over my education. I don't think I could have managed the stress of the PhD student's life without them.

The person I have to thank the most, however, is my loving wife, Hanh Nguyen. She has stood by me through all the craziness, all the trips to Florida, all the drives to California, and everything else while providing an unyielding pillar of support. Even during these last months of my work, when I was staying late in the lab every night to get all my experiments run and my dissertation written, she was handling everything at home while she had an equally busy schedule. Together, we eagerly await the arrival of our daughter, Abigail. There are no words to express how much I appreciate all she have done for me. I am so grateful for all of her love, patience, and support and look forward to starting the next phase of our lives together.

Last but not least, I extend my heartfelt gratitude to my former advisors, Prof. Marios Savvides and Prof. Khoa Luu, as well as my esteemed colleagues: Dr. Zhiqiang Shen, Dr. Chenchen Zhu, Dr. Yutong Chen, Dr. Uzair, Dr. Dipan Pal, Dr. Duong Chi Nhan, Dr. Quach Kha Gia, Dr. Zechun Liu, Ran, Askhay, Nancy, and Devesh at Carnegie Mellon University. Their unwavering support and mentorship laid a robust foundation for my academic research at the outset of my PhD journey.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Identifying the person in a single photo remains challenging because, in many cases, the problem is a zero-shot and ill-posed image retrieval task.

1.1 Motivations

Face identification (FI), the technology for identifying a person from a photo, is ubiquitous and driving many high-stake decisions made by the law enforcement in the United States (Fig. 1.1a). For example, the FI technology has been used to identify attackers of the Jan 2021 US Capitol riot, [34], find shoplifting suspects [67], determine someone is eligible for unemployment benefits [69], and identify ticketed passengers to board an airplane [29]. Yet, the technology can make mistakes, leading to severe consequences, e.g. people wrongly denied of unemployment benefits [69] or falsely arrested [67, 37, 33, 12] (Fig. 1.1b). Face identification is challenging because of several reasons. First, a deep neural network (DNNs) may not have seen a normal, non-celebrity person before during its training. Second, there may be too few photos of a person in the database for FI systems to make reliable decisions. Third, it is harder to identify when a face in the wild (e.g. from surveillance cameras) is occluded [79, 65] (e.g. wearing masks), distant or cropped, yielding a new type of photo not in both the training set of deep networks and the retrieval database—i.e., out-of-distribution (OOD) data. Face verification accuracy notoriously drops significantly (from 99.38% to 81.12% on LFW) given a masked face [65]. This is particularly alarming for FI deployment during this COVID-19 era where masks are mandatory or recommended in public.

In addition, with growing data volumes, fast and high FI systems are paramount for processing and analyzing real-time data to identify faces and patterns effectively. Unfortunately,

(a) Face identification systems are being used for surveillance and assisting law enforcement.

(b) False arrest of an innocent man (right) due to AI errors.

Figure 1.1: Use cases and current problems in current face recognition system [1].

facial information may not always be obtained in ideal conditions, and out-of-distribution data (OOD) e.g. faces with masks, sunglasses, or other adversarial components, poses challenges to correctly identifying the targets. FI accuracy may drop substantially on OOD data, e.g., from 98.41% to 39.79% on LFW when the query face is wearing masks [63] or adversarially modified [103, 4].

Besides the accuracy of FI for OOD data, the field faces two practical challenges. The first challenge is the rapid identification of faces under OOD settings. Swift identification can improve user experience by reducing waiting time during unlocking devices, accessing accounts [30], and security checks [3], increasing people's trust towards machine-generated results [28], and lowering emergency response [53]. The second challenge is how to explain FI decisions to the end-users, which is interestingly understudied. In reality, FI systems are often operated by end-users [64] who expect to get real-time answers and the reasons why such answers are given. The current limited machine-user interoperability causes numerous false decisions [67, 37, 33, 12]. Specifically, only a few studies have produced explanations for FI predictions [63, 77] and none have evaluated the explanations from interpretable FI models on users.

1.2 Contributions

In this study, we devise two novel architectures: DeepFace-EMD [63] and Face-ViT [62] which not only accelerate the computation but also offer an interpretable module for face recognition system

The main contributions of DeepFace-EMD are summarized as follows:

- We propose to evaluate the performance of state-of-the-art facial feature extractors (ArcFace [22], CosFace [89], and FaceNet [70]) on OOD face *identification* tests. That is, our main task is to recognize the person in a query image given a gallery of known faces. Besides in-distribution (ID) query images, we also test FI models on OOD queries that contain (1) common occlusions, i.e. random crops, faces with masks or sunglasses; and (2) adversarial perturbations [103].

- Interestingly, the OOD accuracy can be substantially improved via a 2-stage approach (see Fig. 4.1): First, identify a set of the most globally-similar faces from the gallery using cosine distance and then, re-rank these shortlisted candidates by comparing them with the query at the patch-embedding level using the Earth Mover's Distance (EMD) [66]

- Across three different models (ArcFace, CosFace, and FaceNet), our re-ranking approach consistently improves the original precision (under all metrics: P@1, R-Precision, and MAP@R) *without finetuning* (Sec. 5.2.2) . That is, interestingly, the spatial features extracted from these models can be leveraged to compare images patch-wise (in addition to image-wise) and further improve FI accuracy.

- On masked images [86], our re-ranking method (no training) rivals the ArcFace models finetuned directly on masked images (Sec. 5.2.2) .

In Face-ViT, we explored the design space of ViTs that enable cross-image attention between two input images for FI. On three important criteria (1) accuracy on in-distribution and OOD

data, (2) computational complexity, and (3) explainability, we compare ViTs, CNNs, and EMD-based patch-wise re-ranking methods and find that:

- With cross-image attention, our 2-image Hybrid-ViT model is an effective re-ranking approach. It outperforms traditional FI models (based on CNNs and 1-image ViTs) on both in-distribution and OOD data.

- Our 2-image Hybrid-ViT performs on par with DeepFace-EMD [63]—a state-of-the-art approach to OOD face identification. In addition, our proposed model is more scalable as shown in

  , i.e. running over 2× faster in practice than DeepFace-EMD, which is slow due to the optimal transport optimization phase .

- In a 21-person human study, the users of Hybrid-ViTs and DeepFace-EMD explanations scored substantially higher than the users of Siamese neural networks (SNNs) in face verification . We are the first to report that visual explanations improve end-user accuracy in face verification.

Chapter 2

Realted Works

Face Identification under Occlusion Partial occlusion presents a significant, ill-posed challenge to face identification as the AI has to rely only on incomplete or noisy facial features to make decisions [65]. Most prior methods propose to improve FI robustness by augmenting the training set of deep feature extractors with partially-occluded faces [83, 86, 59, 91, 32, 91]. Training on augmented, occluded data encourages models to rely more on local, discriminative facial features [59]; however, does not prevent FI models from misbehaving on new OOD occlusion types, especially under adversarial scenarios [73]. In contrast, our approach (1) does not require re-training or data augmentation; and (2) harnesses both image-level features (stage 1) and local, patch-level features (stage 2) for FI.

A common alternative is to learn to generate a spatial feature mask [76, 85, 65, 54] or an attention map [91] to *exclude* the occluded (i.e. uninformative or noisy) regions in the input image from the face matching process. Motivated by these works, we tested five methods for inferring the importance of each image patch (Sec. 5.2.1) for EMD computation. Early works used hand-crafted features and obtained limited accuracy [54, 58, 48]. In contrast, the latter attempts took advantage of deep architectures but requires a separate occlusion detector [76] or a masking subnetwork in a custom architecture trained end-to-end [65, 85]. In contrast, we leverage directly the pre-trained state-of-the-art image embeddings (of ArcFace, CosFace, & FaceNet) and EMD to exclude the occluded regions from an input image *without any* architectural modifications or re-training.

Another approach is to predict occluded pixels and then perform FI on the recovered images [99, 90, 106, 92, 36, 47]. Yet, how to recover a non-occluded face while preserving true identity remains a challenge to state-of-the-art GAN-based de-occlusion methods [25, 13, 31].

Re-ranking in Face Identification Re-ranking is a popular 2-stage method for refining image retrieval results [98] in many domains, e.g. person re-identification [68], localization [74], or web image search [20]. In FI, Zhou et. al. [105] used hand-crafted patch-level features to encode an image for ranking and then used multiple reference images in the database to re-rank each top-k candidate. The social context between two identities has also been found to be useful in re-ranking photo-tagging results [10]. Swearingen et. al. [80] found that harnessing an external "disambiguator" network trained to separate a query from lookalikes is an effective re-ranking method. In contrast to the prior work, we do not use extra images [105] or external knowledge [10]. Compared to face re-ranking [26, 60], our method is the first re-rank candidates based on a pair-wise similarity score computed from both the image-level and patch-level similarity computed off of state-of-the-art deep facial features.

EMD for Image Retrieval While EMD is a well-known metric in image retrieval [66], its applications on *deep* convolutional features of images have been relatively under-explored. Zhang et al. [95, 94] recently found that classifying fine-grained images (of dogs, birds, and cars) by comparing them patch-wise using EMD in a deep feature space improves few-shot fine-grained classification accuracy. Yet, their success has been limited to *few*-shot, 5-way and 10-way classification with smaller networks (ResNet-12 [35]). In contrast, here, we demonstrate a substantial improvement in FI using EMD without re-training the feature extractors.

Concurrent to our work, Zhao et al. [101] proposes DIML, which exhibits consistent improvement of $\sim$2–3% in image retrieval on images of birds, cars, and products by using the sum of cosine distance and EMD as a "structural similarity" score for ranking. They found that CC is more effective than assigning uniform weights to image patches [99]. Interestingly, via a rigorous study into different feature-weighting techniques, we find novel insights specific for FI: Uniform weighting is more effective than CC. Unlike prior EMD works [101, 95, 94, 87], ours is **the first to** show the significant effectiveness of EMD on (1) occluded and adversarial OOD images; and (2) on face identification.

Out-of-distribution face identification. Identifying faces under occlusion [63, 86, 65] or adversarial changes [103] is challenging. FI systems using SNNs are vulnerable to images containing sunglasses, masks, or adversarial perturbations. A line of approach re-trains deep CNN feature extractors on images with partially-occluded faces [83, 86, 59, 91, 32, 91]. However, data augmentation on a specific type of occlusion (e.g. face masks) does not guarantee generalization to new OOD changes (e.g. in hairstyles) in the input image [63]. An alternative technique for OOD face data is to reconstruct the missing pixels before performing FI [90, 106, 92, 36, 47, 100]. Yet, the de-occlusion process [25, 13, 31] may fail to preserve the identity of the target person and add another level of abstraction over how the FI system computes its decisions, further opaquing the decision-making process.

Siamese networks for patch-wise comparison. A common FI technique involves adopting the Siamese architecture, feeding a pair of input images into two weight-shared, CNN-based feature extractors, and comparing the cosine similarity between two output image-level embeddings [50, 22, 70, 89]. Recent EMD-based image similarity work found that combining both image-level and patch-level similarity yields higher accuracy on in-distribution data [101] and OOD data [96, 63]. DeepFace-EMD [63] consistently outperforms traditional methods [89, 22, 70] that are based on the cosine similarity of two image embeddings from a SNN. Such approaches, however, only conduct a global, image-level comparison and may discard useful local, patch-level information. Researchers are looking for more accurate and efficient architectures for FI tasks.

Vision Transformers for patch-wise comparison. Operating at the patch level, ViTs are increasingly popular in computer vision [27, 41, 82, 107], were shown to achieve remarkable image classification accuracy, and do not need explicit feature extraction like in CNN-based models. Most ViT research focuses on a *single*-image architecture where self-attention [84] is leveraged to compare the similarity between *intra-image* patches [104] or between image-patches and text-tokens in image-text architectures [38, 46]. CrossViT [18] proposed to use two Transformers but for two differently scaled versions of *the same* image,

not for two images. The only work utilizing ViTs in FI that we are aware of is the concurrent work by [104], which uses the vanilla ViT on a *single* image and therefore offers no cross-image interaction. A few other concurrent works also explore ViTs for 2-image inputs but rather for person re-identification [88, 49], a different task that involves a more unconstrained image distribution than the images typically cropped and aligned in FI. These leave us great room for exploring cross-image interaction to compare two face images.

Model interpretability of Vision Transformers. Various efforts have been made to visualize the effects of ViTs. Black et al. [11] proposed a novel method to combine cross-correlation and an attention flow approximation between two images, each processed by a different 1-image ViT. For multimodal, vision-language Transformers, Kim et al. [38] use the similarity flow between text and image tokens as explanations for its similarity score. Chefer et al. [16, 17] leveraged the aggregate cross-attention across layers and its gradients to derive a visualization of similarity between two inputs. In our work, we visualize all ViTs using the technique proposed by [77].

Chapter 3

DeepFace-EMD: Re-ranking Using Patch-wise Earth Mover's Distance Improves Out-Of-Distribution Face Identification

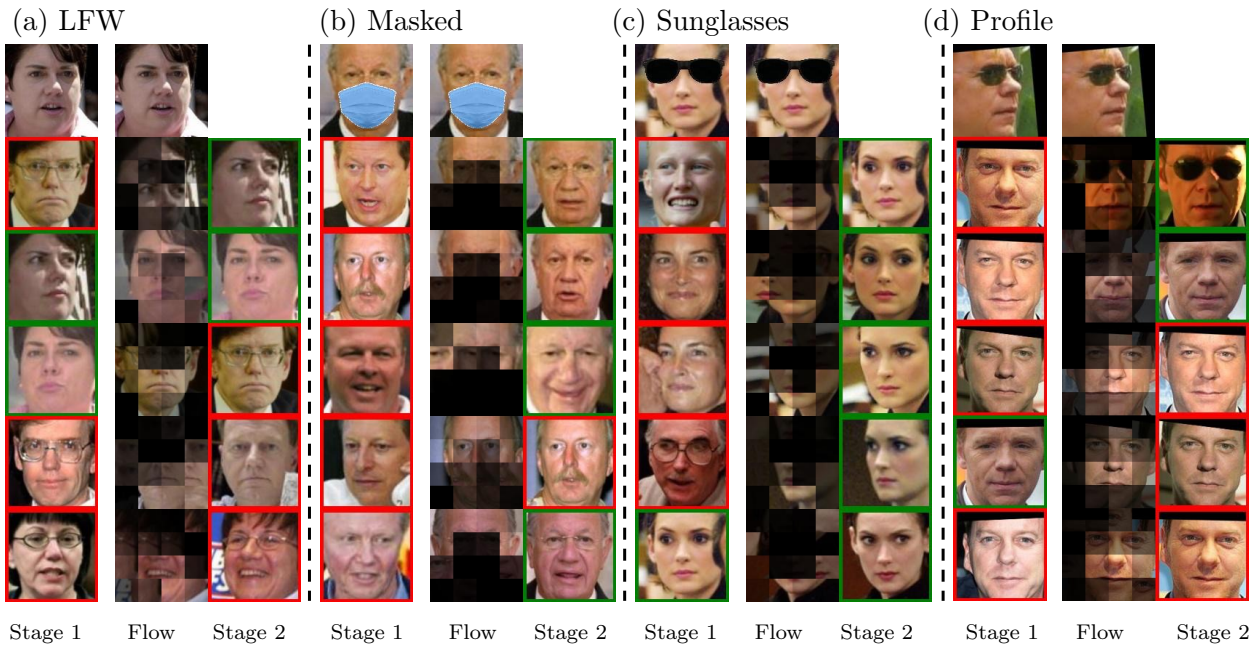


Figure 3.1: Traditional face identification ranks gallery images based on their cosine distance with the query (top row) at the image-level embedding, which yields large errors upon out-of-distribution changes in the input (e.g. masks or sunglasses; b–d). We find that re-ranking the top-k shortlisted faces from Stage 1 (leftmost column) using their patch-wise EMD similarity w.r.t. the query substantially improves the precision (Stage 2) on challenging cases (b–d). The "Flow" visualization intuitively shows the patch-wise reconstruction of the query face using the most similar patches (i.e. highest flow) from the retrieved face.

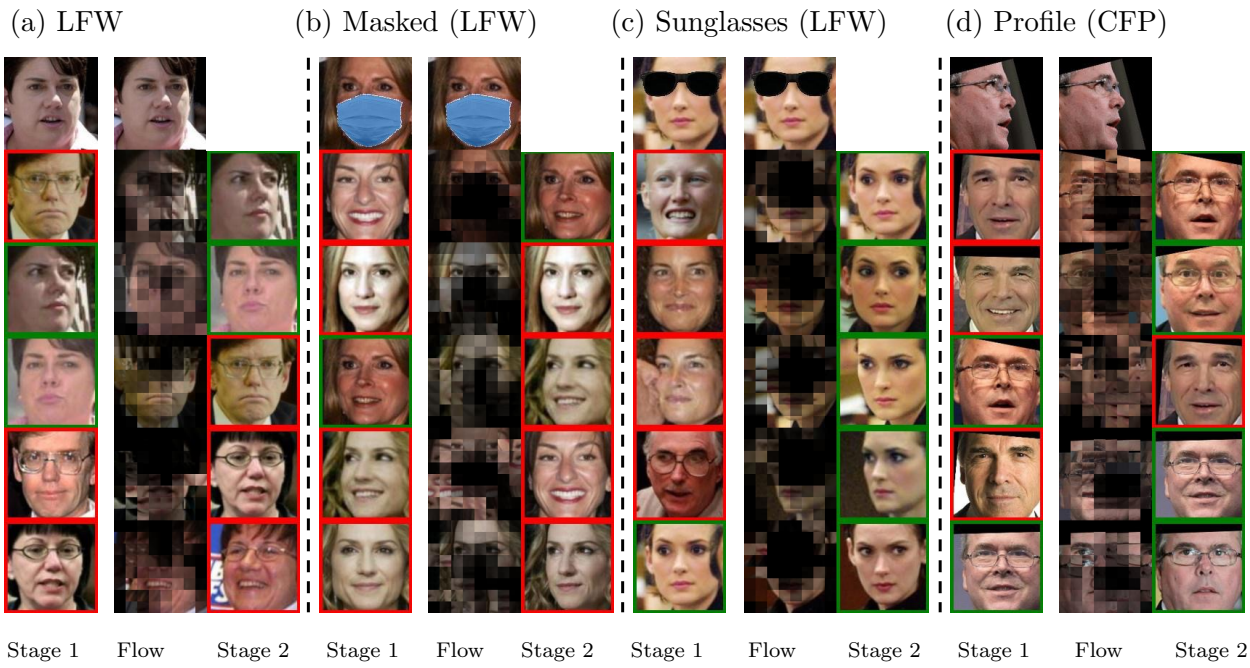


Figure 3.2: Traditional face identification ranks gallery images based on their cosine distance with the query (top row) at the image-level embedding, which yields large errors upon out-of-distribution changes in the input (e.g. masks or sunglasses; b–d). We find that re-ranking the top-k shortlisted faces from Stage 1 (leftmost column) using their patch-wise EMD similarity w.r.t. the query substantially improves the precision (Stage 2) on challenging cases (b–d). The "flow" visualization (of 8×8) intuitively shows the patch-wise reconstruction of the query face using the most similar patches (i.e. highest flow) from the retrieved face.

3.1 Networks

3.1.1 Pre-trained models

We use three state-of-the-art PyTorch models of ArcFace, FaceNet, and CosFace pre-trained on CASIA [93], VGGFace2 [14], and CASIA, respectively. Their architectures are ResNet-18 [35], Inception-ResNet-v1 [81], and 20-layer SphereFace [50], respectively. For more details on network architectures and implementation in PyTorch.

We downloaded the three pre-trained PyTorch models of ArcFace, FaceNet, and CosFace from:

- ArcFace [22]: `https://github.com/ronghuaiyang/arcface-pytorch`
- FaceNet [70]: `https://github.com/timesler/facenet-pytorch`

- CosFace [89]: `https://github.com/MuggleWang/CosFace_pytorch`

These ArcFace, FaceNet, and CosFace models were trained on dataset CASIA Webface [93], VGGFace2 [15], and CASIA Webface [93], respectively.

The network architectures are provided here:

- ArcFace: `https://github.com/ronghuaiyang/arcface-pytorch/blob/master/models/resnet.py`
- FaceNet: `https://github.com/timesler/facenet-pytorch/blob/master/models/inception_resnet_v1.py`
- CosFace: `https://github.com/MuggleWang/CosFace_pytorch/blob/master/net.py#L19`

3.1.2 Image pre-processing

For all networks, we align and crop input images following the 3D facial alignment in [9] (which uses 5 reference points, 0.7 and 0.6 crop ratios for width and height, and Similarity transformation). All images shown in this paper (e.g. Fig. 3.1) are pre-processed. Using MTCNN, the default pre-processing of all three networks, does not change the results substantially (See Sec. 5.2.1).

3.1.3 2-stage hierarchical face identification

Stage-1: Ranking A common 1-stage face identification [50, 70, 89] ranks gallery images based on their pair-wise cosine similarity with a given query in the last-linear-layer feature space of a pre-trained feature extractor (Fig. 4.1). Here, our image embeddings are extracted from the *last linear* layer of all three models and are all $\in \mathbb{R}^{512}$.

Stage-2: Re-ranking We re-rank the top-k (where the optimal $k = 100$) candidates from Stage 1 by computing the patch-wise similarity for an image pair using EMD. Overall, we

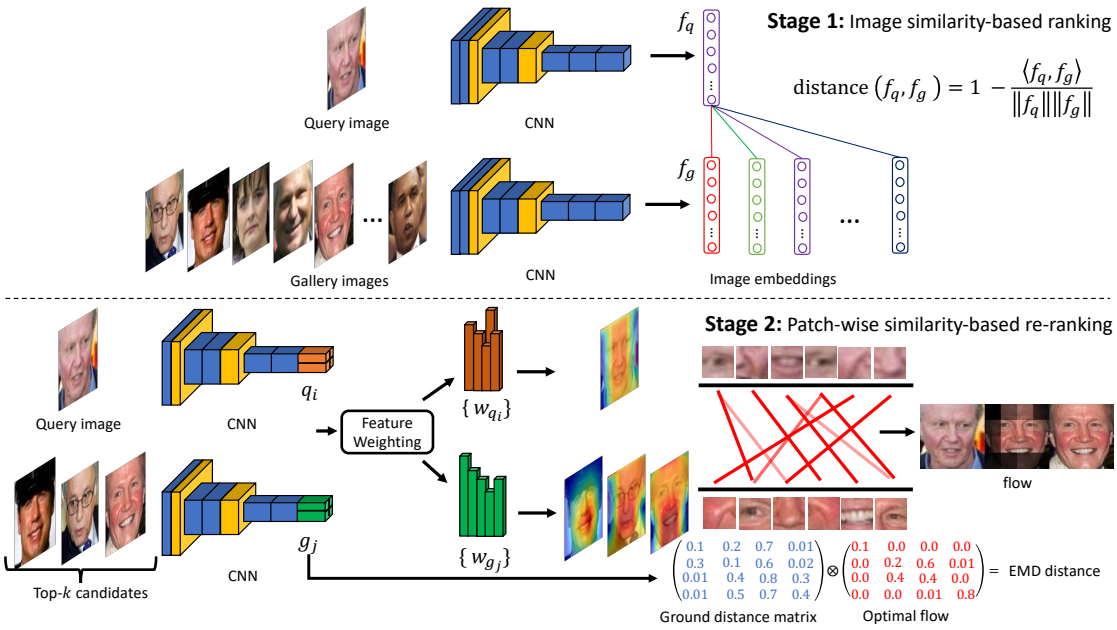


Figure 3.3: Our 2-stage face identification pipeline. Stage 1 ranks gallery images based on their cosine distance with the query face at the image-embedding level. Stage 2 then re-ranks the top-k shortlisted candidates from Stage 1 using EMD at the patch-embedding level.

compare faces in **two hierarchical stages** (Fig. 4.1), first at a coarse, image level and then a fine-grained, patch level.

Via an ablation study (Sec. 5.2.1) , we find our 2-stage approach (a.k.a. DeepFace-EMD) more accurate than Stage 1 alone (i.e. no patch-wise re-ranking) and also Stage 2 alone (i.e. sorting the entire gallery using patch-wise similarity).

3.1.4 Earth Mover's Distance (EMD)

EMD is an edit distance between two set of weighted objects or distributions [66]. Its effectiveness was first demonstrated in measuring pair-wise image similarity based on color histograms and texture frequencies [66] for image retrieval. Yet, EMD is also an effective distance between two text documents [44], probability distributions (where EMD is equivalent to Wasserstein, i.e. Mallows distance) [45], and distributions in many other domains [61, 51, 42]. Here, we propose to harness EMD as a distance between two faces, i.e. two sets of weighted facial features.

Let $\mathcal{Q} = \{(q_1, w_{q_1}), ..., (q_N, w_{q_N})\}$ be a set of N (facial feature, weight) pairs describing a query face where q_i is a feature (e.g. left eye or nose) and the corresponding w_{q_i} indicates how important the feature q_i is in FI. The *flow* between $\mathcal{Q}$ and the set of weighted features of a gallery face $\mathcal{G} = \{(g_1, w_{g_1}), ..., (g_N, w_{g_N})\}$ is any matrix $\boldsymbol{F} = (f_{ij}) \in \mathbb{R}^{N \times N}$. Intuitively, f_{ij} is the amount of importance weight at q_i that is matched to the weight at g_j. Let d_{ij} be a ground distance between (q_i, g_j) and $\boldsymbol{D} = (d_{ij}) \in \mathbb{R}^{N \times N}$ be the ground distance matrix of all pair-wise distances.

We want to find an optimal flow $\boldsymbol{F}$ that minimizes the following cost function, i.e. the sum of weighted pair-wise distances across the two sets of facial features:

$$\text{COST}(\mathcal{Q}, \mathcal{G}, \boldsymbol{F}) = \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} f_{ij} \tag{3.1}$$

$$\text{s.t.} \qquad f_{ij} \geq 0 \tag{3.2}$$

$$\sum_{j=1}^{N} f_{ij} \leq w_{q_i}, \text{and} \sum_{i=1}^{N} f_{ij} \leq w_{g_j}, i, j \in [1, N] \tag{3.3}$$

$$\sum_{j=1}^{N} \sum_{i=1}^{N} f_{ij} = \min \left(\sum_{j=1}^{N} w_{g_j}, \sum_{i=1}^{N} w_{q_i} \right). \tag{3.4}$$

As in [101, 94], we normalize the weights of a face such that the total weights of features is 1 i.e. $\sum_{i=1}^{N} w_{q_i} = \sum_{j=1}^{N} w_{g_j} = 1$, which is also the total flow in Eq. (3.4). Note that EMD is a metric iff two distributions have an equal total weight and the ground distance function is a metric [19].

We use the iterative Sinkhorn algorithm [21] to efficiently solve the linear programming problem in Eq. (3.1), which yields the final EMD between two faces $\mathcal{Q}$ and $\mathcal{G}$.

Facial features In image retrieval using EMD, a set of features $\{q_i\}$ can be a collection of dominant colors [66], spatial frequencies [66], or a histogram-like descriptor based on the local patches of reference identities [87]. Inspired by [87], we also divide an image into a grid but we take the embeddings of the local patches from the last convolutional layers of each network. That is, in FI, face images are aligned and cropped such that the entire face covers

most of the image (see Fig. 3.1a). Therefore, without facial occlusion, every image patch is supposed to contain useful identity information, which is in contrast to natural photos [94].

Our grid sizes $H \times W$ for ArcFace, FaceNet, and CosFace are respectively, 8×8, 3×3, and 6×7, which are the corresponding spatial dimensions of their last convolutional layers. That is, each feature q_i is an embedding of size $1 \times 1 \times C$ where C is the number of channels (i.e. 512, 1792, and 512 for ArcFace, FaceNet, and CosFace, respectively).

Ground distance Like [94, 101], we use cosine distance as the ground distance d_{ij} between the embeddings (q_i, g_j) of two patches:

$$d_{ij} = 1 - \frac{\langle q_i, g_j \rangle}{\|q_i\| \, \|g_j\|} \tag{3.5}$$

where $\langle . \rangle$ is the dot product between two feature vectors.

3.1.5 Feature weighting

EMD in our FI intuitively is an optimal plan to match all weighted features across two images. Therefore, how to weight features is an important step. Here, we thoroughly explore five different feature-weighting techniques for FI.

Uniform Zhang et al. [94] found that it is beneficial to assign lower weight to less informative regions (e.g. background or occlusion) and higher weight to discriminative areas (e.g. those containing salient objects). Yet, assigning an equal weight to all $N = H \times W$ patches is worth testing given that background noise is often cropped out of the pre-processed face image (Fig. 3.1):

$$w_{q_i} = w_{g_i} = \frac{1}{N}, \text{where } 1 \leq k \leq N \tag{3.6}$$

Average Pooling Correlation (APC) Instead of uniformly weighting all patch embeddings, an alternative from [94] would be to weight a given feature q_i proportional to its correlation to the entire other *image* in consideration. That is, the weight w_{q_i} would be the

dot product between the feature q_i and the average pooling output of all embeddings $\{g_j\}_1^N$ of the gallery image:

$$w_{q_i} = \max\big(0, \langle q_i, \frac{\sum_j^N g_j}{N}\rangle\big), w_{g_j} = \max\big(0, \langle g_j, \frac{\sum_i^N q_i}{N}\rangle\big) \tag{3.7}$$

where max(.) keeps the weights always non-negative. APC tends to assign near-zero weight to occluded regions and, interestingly, also minimizes the weight of eyes and mouth in a non-occluded gallery image (see Fig. 3.4b; blue shades around both the mask and the non-occluded mouth).

Cross Correlation (CC) APC [94] is different from CC introduced in [101], which is the same as APC except that CC uses the output vector from the last linear layer (see code) instead of the global average pooling vector in APC.

Spatial Correlation (SC) While both APC and CC "summarize" an entire other gallery image into a vector first, and then compute its correlation with a given patch q_i in the query. In contrast, an alternative, inspired by [78], is to take the sum of the cosine similarity between the query patch q_i and every patch in each gallery image $\{g_j\}_1^N$:

$$w_{q_i} = \max\big(0, \sum_j^N \frac{\langle q_i, g_j\rangle}{\|q_i\|\|g_j\|}\big), w_{g_j} = \max\big(0, \sum_i^N \frac{\langle q_i, g_j\rangle}{\|q_i\|\|g_j\|}\big) \tag{3.8}$$

We observe that SC often assigns a higher weight to occluded regions e.g., masks and sunglasses (Fig. 3.4b).

Landmarking (LMK) While the previous three techniques adaptively rely on the image-patch similarity (APC, CC) or patch-wise similarity (SC) to weight a given patch embedding, their considered important points may or may not align with facial landmarks, which are known to be important for many face-related tasks. Here, as a baseline for APC, CC, and SC, we use dlib [40] to predict 68 keypoints in each face image (see Fig. 3.4c) and weight each

patch-embedding by the density of the keypoints inside the patch area. Our LMK weight distribution appears Gaussian-like with the peak often right below the nose (Fig. 3.4c).

3.2 Flow visualization

We use the same visualization technique as in DeepEMD to generate the flow visualization showing the correspondence between two images (see the flow visualization in Fig. 3.1 or Fig. 3.4). Given a pair of embeddings from query and gallery images, EMD computes the optimal flows (see Eq. (3.1) for details). That is, given a 8×8 grid, a given patch embedding q_i in the query has 64 flow values $\{f_{ij}\}$ where $j \in \{1, 2, ..., 64\}$. In the location of patch q_i in the query image, we show the corresponding highest-flow patch g_k, i.e. k is the index of the gallery patch of highest flow $f_{i,k} = \max(f_{i,1}, f_{i,2}, ..., f_{i,64})$. For displaying, we normalize a flow value $f_{i,k}$ over all 64 flow values (each for a patch $i \in \{1, 2, ..., 64\}$) via:

$$f = \frac{f - \min(f)}{\max(f) - \min(f)} \tag{3.9}$$

See Fig. 3.1, Fig. 3.2, and Fig. 5.2 for example flow visualizations.

(a) SC (b) APC (c) LMK (d) Uniform

Figure 3.4: Given a pair of images, after the features are weighted (heatmaps; red corresponds to 1 and blue corresponds to 0 importance weight), EMD computes an optimal matching or "transport" plan. The middle flow image shows the one-to-one correspondence following the format in [94] (see also description in 3.2). That is, intuitively, the flow visualization shows the reconstruction of the left image, using the nearest patches (i.e. highest flow) from the right image. Here, we use ArcFace and a 4× 4 patch size (i.e. computing the EMD between two sets of 16 patch-embeddings). Darker patches correspond to smaller flow values. How EMD computes facial patch-wise similarity differs across different feature weighting techniques (SC, APC, LMK, and Uniform). Based on per-patch density of detected landmarks (- - -), LMK (c) often assigns higher weight to the center of a face (regardless of occlusions).

Chapter 4

Fast and Interpretable Face Recognition for Out-Of-Distribution Data Using Vision Transformers (ViTs)

We propose a novel ViT architecture (denoted as Model H2L) for FI on OOD data. It takes in *two* images as input to leverage both self-attention and cross-attention to compute a similarity score for two images.

4.1 Problem Formulation

Similar to DeepFace-EMD [63], our method identifies a person in a query image by ranking all gallery images based on their pair-wise similarity with the query. After ranking (ST1) or re-ranking (ST2), we take the top-1 nearest image as the predicted identity. For the scope of this paper, we only consider data consisting of frontal faces without gestures.

4.2 Architecture: a two-Image Hybrid ViT

The overall architecture of the model is shown in Tab. 4.1. and Fig. 4.1. It takes in patch embeddings from a pre-trained CNN (ArcFace [22]). The Transformer encoder consists of a block of a multiheaded self-attention (MSA) layer and an MLP layer. After N layers of the Transformer encoder, which contains both self-attention and cross-attention from 2 input images, the patch embeddings of the input images go through two linear layers.

Face embeddings. For a zero-shot face problem, deep metric learning works efficiently [70, 50, 89]. Besides *[CLS]* (classification tokens) [23, 27] for feature embeddings, we also use the remaining 2-output to separate linear layers to extract features that are deployed to a deep metric learning fashion (see Fig. 4.1 for details).

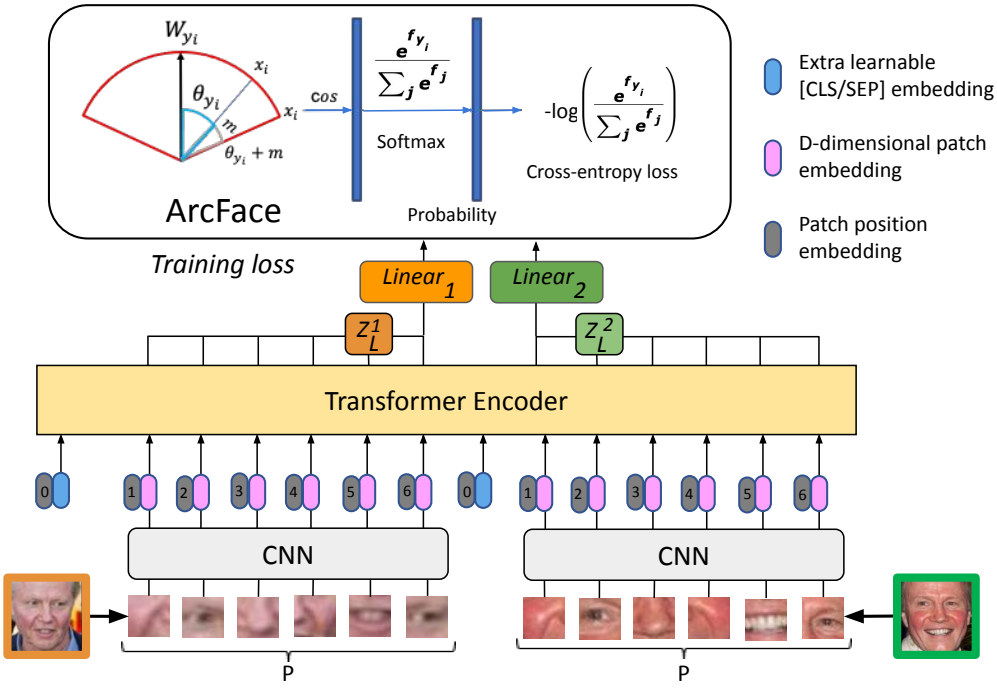


Figure 4.1: The architecture of the proposed ViT-based Model H2L.

Given two input 2D face images $\mathbf{x}_1, \mathbf{x}_2$, we reshape them to have dimensions $\in \mathbb{R}^{H\times W\times C}$. The face embeddings $\mathbf{x}_{p1}, \mathbf{x}_{p2} \in \mathbb{R}^{P^2\times D}$ are extracted from either CNNs or a linear embedding layer, where P is the number of the face patches and D is the size of the patch embedding. Here in the loss function ArcFace [22], we use $D = 512$ and $P = 8$.

We denote the learnable embeddings as $\mathbf{E}$ and $\mathbf{E}_{pos} \in \mathbb{R}^{(2\times P^2+2)\times D}$, the two extra learnable embeddings as $\mathbf{X}_{CLS}$ and $\mathbf{X}_{SEP}$, and the intermediate layers of the Transformer encoder as $\mathbf{z}_i$. $\mathbf{f}_1$ and $\mathbf{f}_2$ are the features from *two* linear layers that contain cross-attention information between two images. Our proposed two-image-based model can be formulated as follows.

$$\mathbf{z}_0 = [\mathbf{x}_{CLS}\mathbf{E}, \mathbf{x}_{p1}\mathbf{E}, \mathbf{x}_{SEP}\mathbf{E}, \mathbf{x}_{p2}\mathbf{E}] + \mathbf{E}_{pos}, \tag{4.1}$$

$$\mathbf{z}_l^{'} = \text{MSA}(\text{LayerNorm}(\mathbf{z}_{l-1})), \quad l = 1 \dots L \tag{4.2}$$

$$\mathbf{z}_l = \text{MLP}(\text{LayerNorm}(\mathbf{z}_l^{'})) + \mathbf{z}_l^{'}, \ l = 1 \dots L \tag{4.3}$$

$$\mathbf{z}_l \equiv [\mathbf{z}_{CLS}, \mathbf{z}_L^1, \mathbf{z}_{SEP}, \mathbf{z}_L^2], \quad \mathbf{z}_L^1, \mathbf{z}_L^2 \in \mathbb{R}^{P^2 \times D} \tag{4.4}$$

$$\mathbf{f}_1 = \text{LayerNorm}(\text{Linear}_1(\mathbf{z}_L^1)) \tag{4.5}$$

$$\mathbf{f}_2 = \text{LayerNorm}(\text{Linear}_2(\mathbf{z}_L^2)) \tag{4.6}$$

$$\text{loss} = \text{Arcface\_loss}(\mathbf{f}_1, \mathbf{f}_2) \tag{4.7}$$

Position embeddings in vanilla Transformers [84] indicate the position of words in sentences for machine translation. Here, they are also used with the face inputs. When parts of the face are arranged in a constrained order, e.g. position of eyes, mouth, etc. this positioning information maintains the facial structure.

Attention-based outputs. The outputs $\mathbf{z}_l^{'}$ from a multi-head-attention (MSA) layer are obtained through a combination of self and cross-attention processes. Previous ViT works [27, 39, 7, 18] usually apply *[CLS]* as an extra learnable embedding for specific tasks. However, similar to spatial patch embeddings in CNNs, the two-image-input-based model exploits the patch embedding output $\mathbf{z}_L^1, \mathbf{z}_L^2$ which contain information from both images, then put them into linear layers for extracting cross-image features. We provide an ablation study to compare the performance of these cross-image features and *[CLS]* in Sec. 5.3.1. Similar to previous deep metric learning methods in face recognition [70, 89, 50], here we use the ArcFace as our loss function [22] to separate and learn cross-image margins to their corresponding labels.

Name	Architecture	Patch Embedding	Input	Transformer output	**Inter**-image, Image-wise comparison	**Intra**-image, patch-wise comparison	**Inter**-image, patch-wise comparison
C	CNN [22]	CNN [2]	1-image	1 feature	✓	Local (CNN-based)	✗
V	ViT [27]	*learned*	1-image	1 feature	✓	✓	✗
H1	Hybrid-ViT	CNN	1-image	1 feature	✓	✓	✗
H2	Hybrid-ViT	CNN	2-image	CLS	✗	✓	✓
H2L	Hybrid-ViT (**ours**)	CNN	2-image	2-Linear	✓	✓	✓
D	DeepFace-EMD [63]	CNN	2-image	2 features	✓($\alpha = 0.3$)	Local (CNN-based)	✓($\alpha = 0.7$)

Table 4.1: Properties of the six networks evaluated in this work. We categorize into 2 types of models: 1-image and 2-image. 1-image models include CNN (C) and ViT (V) while the 2-image group contains DeepFace-EMD (D). Hybrid-ViT can be 1-image (H1) or 2-image (H2 and H2L). The difference between H2 and H2L is the Transformer output of *[CLS]* vs. 2-Linear, respectively.

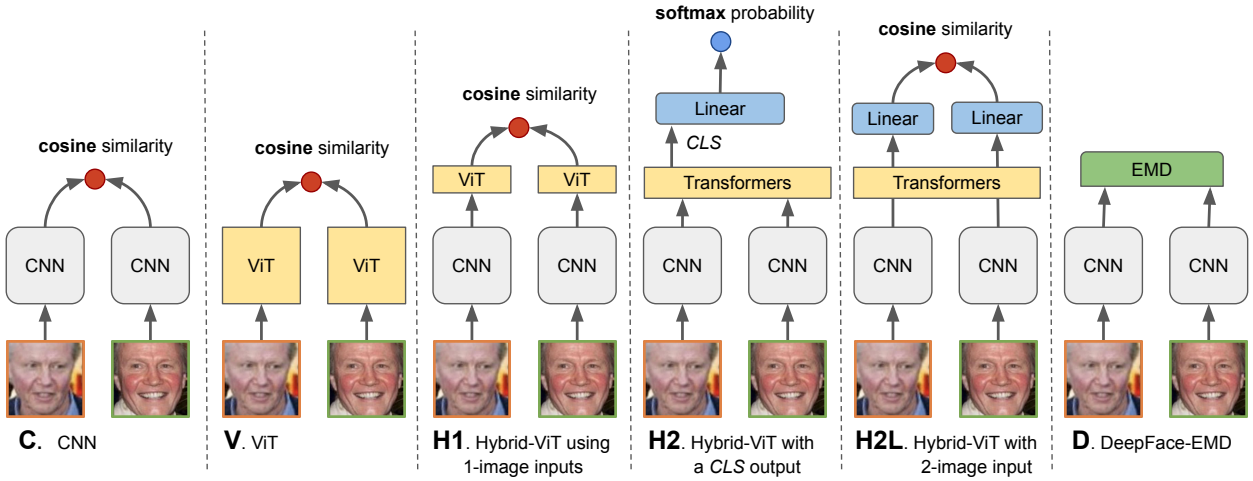


Figure 4.2: The architecture of the six networks evaluated in this work including our proposed H2L.

4.3 Dataset

The model is trained on the CASIA Webface [93] dataset, containing 494,414 face images of 10,575 real-world identities, widely used for FI tasks such as [70]. We sample 2M pairs (1M positives and 1M negatives) consisting of all identities from the processed and clean CASIA Webface dataset.

4.4 Evaluation against various network structures

Here, we study six models with various architectures for face recognition, including a SNN with ArcFace [22], DeepFace-EMD [63], and Transformer ViTs, whose properties are summarized in Tab. 4.1.

The Siamese CNN model (denoted as C in the table) is used as a baseline in our study. The ViT-based model (denoted as V) operates at the patch level instead of the image level. The 1-image hybrid-ViT [27] (Model H1) is the same as the original ViT except that the patch embeddings are from a pre-trained CNN, which serves as the baseline for ViT-based models. The 2-image Hybrid-ViT (Model H2) uses [CLS] for binary cross-entropy loss for one single softmax classifier layer, which we will compare to the 1-image model. The 2-image Hybrid-ViT (Model H2L) uses 2-output features for computing a cosine similarity. The 1-image model has separate ViTs for each input while the 2-image one has put two features into a single Transformer to implement cross-attention. DeepFace-EMD [63] (D) uses entire CNN features but in two stages: First, compare images using image embeddings and then re-rank using patch embeddings. Models H2, H2L, & D perform cross-image, patch-wise comparison—via ViT attention (H2 & H2L) or optimal transport (D) between 2 image inputs.

For Model H2L, the spatial features embeddings (e.g. 8×8 in ResNet-18 [35]) are re-used to compute a feature vector through the linear layers which are deployed to ArcFace [22] loss function. Utilizing this loss function for cross-image features can help transfer knowledge quickly as well as further improvements. For more details about parameter selection, see Tab. 4.1 and Sec. 4.4.

Chapter 5

Experimental Results

5.1 Evaluation metrics

P@1 is well-known as Recall@1 in metric learning. P@1 is computed as follow.

$$\mathcal{N}_q^k = \underset{\mathcal{N}\subset\mathcal{X}_{\text{test}}\ ,|\mathcal{N}|=k}{\arg\min} \sum_{x^f\in\mathcal{N}} d_e\left(\phi\left(x^q\right),\phi\left(x^f\right)\right)$$

where x^q and $\phi(\cdot)$ are inputs and feature encoder respectively, $d_e(\cdot,\cdot)$ is the euclidean distance, and k is k-nearest neighbors. Precision@k can be calculated as:

$$\text{P@}k = \frac{1}{|\mathcal{X}_{\text{test}}|}\sum_{x_q\in\mathcal{X}_{\text{test}}}\frac{1}{k}\sum_{x^i\in\mathcal{N}_q^k}\begin{cases}1, & y^i = y^q\\ 0, & \text{otherwise}\end{cases}$$

where y^i is the class label of sample x^i.

To gain more information and a comprehensive ranking evaluation, we computed mean average precision of R (M@R [56]), where R is number of images in a class.

$$M@R = \frac{1}{R}\sum_{i=1}^{R} P(i)$$

where

$$P(i) = \begin{cases}\text{P@}i, & \text{if the } i\text{-th retrieval is correct;}\\ 0, & \text{otherwise.}\end{cases}$$

5.2 DeepFace-EMD: Re-ranking Using Patch-wise Earth Mover's Distance Improves Out-Of-Distribution Face Identification

5.2.1 Ablation Studies

We perform three ablation studies to rigorously evaluate the key design choices in our 2-stage FI approach: (1) Which feature-weighting techniques to use (Sec. 5.2.1)? (2) re-ranking using both EMD and cosine distance (Sec. 5.2.1); and (3) comparing patches or images in Stage 1 (Sec. 5.2.1).

Experiment For all three experiments, we use ArcFace to perform FI on both LFW [93] and LFW-crop. For **LFW**, we take all 1,680 people who have $\geq$2 images for a total of 9,164 images. When taking each image as a query, we search in a gallery of the remaining 9,163 images. For the experiments with **LFW-crop**, we use all 13,233 original LFW images as the gallery. To create a query set of 13,233 cropped images, we clone the gallery and crop each image randomly to its 70% and upsample it back to the original size of 128×128 (see examples in Fig. 5.2d). That is, LFW-crop tests identifying a cropped (i.e. close-up, and misaligned) image given the unchanged LFW gallery. LFW and LFW-crop tests offer contrast insights (ID vs. OOD).

In Stage 2, i.e. re-ranking the top-k candidates, we test different values of $k \in \{100, 200, 300\}$ and do not find the performance to change substantially. At $k = 100$, our 2-stage precision is already close to the maximum precision of 99.88 under a perfect re-ranking (see Tab. 5.2a; Max prec.).

3D Facial Alignment vs. MTCNN

The reason we used the 3D alignment pre-processing [9] instead of the default MTCNN pre-processing [97] of the three models was because for ArcFace, the 3D alignment actually resulted in better P@1, RP, and M@R for both our baselines and DeepFace-EMD (e.g. +3.35% on MLFW). For FaceNet, the 3D alignment did yield worse performance compared

to MTCNN. However, we confirm that our conclusions that **DeepFace-EMD improves FI on the reported datasets regardless of the pre-processing choice**. See Tab. 5.1 for details.

Dataset	Model	Pre-processing	Method	P@1	RP	M@R
CALFW (Mask)	ArcFace	3D alignment	ST1	96.81	53.13	51.70
			Ours	**99.92**	**57.27**	**56.33**
		MTCNN	ST1	96.36	48.35	46.85
			Ours	**99.92**	**53.53**	**52.53**
	FaceNet	3D alignment	ST1	77.63	39.74	36.93
			Ours	**96.67**	**45.87**	**44.53**
		MTCNN	ST1	86.65	45.29	42.83
			Ours	**98.62**	**49.75**	**48.49**
AgeDB (Mask)	ArcFace	3D alignment	ST1	96.15	39.22	30.41
			Ours	**99.84**	39.22	**33.18**
		MTCNN	ST1	95.35	29.51	22.75
			Ours	**99.78**	**32.82**	**27.08**
	FaceNet	3D alignment	ST1	75.99	22.28	14.95
			Ours	**96.53**	**24.25**	**17.49**
		MTCNN	ST1	83.93	25.18	17.74
			Ours	**98.26**	**27.27**	**20.45**

Table 5.1: DeepFace-EMD improved FI on the reported datasets regardless of the pre-processing choice.

Comparing feature weighting techniques

Here, we evaluate the precision of our 2-stage FI as we sweep across five different feature-weighting techniques and two grid sizes (8×8 and 4×4). In an 8×8 grid, we observe that some facial features such as the eyes are often split in half across two patches (see Fig. 3.2), which may impair the patch-wise similarity. Therefore, for each weighting technique, we also test average-pooling the 8×8 grid into 4×4 and performing EMD on the resultant 16 patches.

Results First, we find that, on LFW, our image-similarity-based techniques (APC, SC) outperform the LMK baseline (Tab. 5.2a) despite not using landmarks in the weighting process, verifying the effectiveness of adaptive, similarity-based weighting schemes.

Second, interestingly, in FI, we find that Uniform, APC, and SC all outperform the CC weighting proposed in [101, 94]. This is in stark contrast to the finding in [101] that CC is better than Uniform (perhaps because face images do not have background noise and are close-up). Furthermore, using the global average-pooling vector from the channel (APC) substantially yields more useful spatial similarity than the last-linear-layer output as in CC implementation (Tab. 5.2b; 96.16 vs. 91.31 P@1).

Third, surprisingly, despite that a patch in a 8×8 grid does not enclose an entire, fully-visible facial feature (e.g. an eye), all feature-weighting methods are on-par or better on an 8×8 grid than on a 4×4 (e.g. Tab. 5.2b; APC: 96.16 vs. 95.32). Note that the optimal flow visualized in a 4×4 grid is more interpretable to humans than that on a 8×8 grid (compare Fig. 3.1 vs. Fig. 3.2).

Fourth, across all variants of feature weighting, our 2-stage approach consistently and *substantially* outperforms the traditional Stage 1 alone on LFW-crop, suggesting its robust effectiveness in handling OOD queries.

Fifth, under a perfect re-ranking of the top-k candidates (where $k = 100$), there is only 1.4% headroom for improvement upon Stage 1 alone in LFW (Tab. 5.2a; 98.48 vs. 99.88) while there is a large ∼12% headroom in LFW-crop (Tab. 5.2a; 87.35 vs. 98.71). Interestingly, our re-ranking results approach the upperbound re-ranking precision (e.g. Tab. 5.2b; 96.26 of Uniform vs. 98.71 Max prec. at $k = 100$).

Re-ranking using both EMD & cosine distance

We observe that for some images, re-ranking using patch-wise similarity at Stage 2 does not help but instead hurt the accuracy. Here, we test whether linearly combining EMD (at the *patch*-level embeddings as in Stage 2) and cosine distance (at the *image*-level embeddings as in Stage 1) may improve *re-ranking* accuracy further (vs. EMD alone).

ArcFace	Method	P@1	RP	M@R
LFW vs. LFW (a)	Stage 1 alone [22]	98.48	78.69	78.29
	Max prec. at $k = 100$	99.88	81.32	-
	CC [101] (8×8)	98.42	78.35	77.91
	CC [101] (4×4)	81.69	76.29	72.47
	APC (8×8)	**98.60**	78.63	78.23
	APC (4×4)	**98.54**	78.57	78.16
	Uniform (8×8)	**98.66**	**78.73**	**78.35**
	Uniform (4×4)	**98.63**	**78.72**	**78.33**
	SC (8×8)	**98.66**	**78.74**	**78.35**
	SC (4×4)	**98.65**	**78.72**	**78.33**
	LMK (8×8)	98.35	78.43	77.99
	LMK (4×4)	98.31	78.38	77.90
LFW-crop vs. LFW (b)	Stage 1 alone [22]	87.35	71.38	69.04
	Max prec. at $k = 100$	98.71	89.13	-
	CC [101] (8×8)	91.31	72.33	70.00
	CC [101] (4×4)	63.12	56.03	51.00
	APC (8×8)	**96.16**	**76.60**	**74.57**
	APC (4×4)	**95.32**	**75.37**	**73.25**
	Uniform (8×8)	**96.26**	**78.08**	**76.25**
	Uniform (4×4)	**95.53**	**77.15**	**75.29**
	SC (8×8)	**96.19**	**78.05**	**76.20**
	SC (4×4)	**95.42**	**77.12**	**75.25**

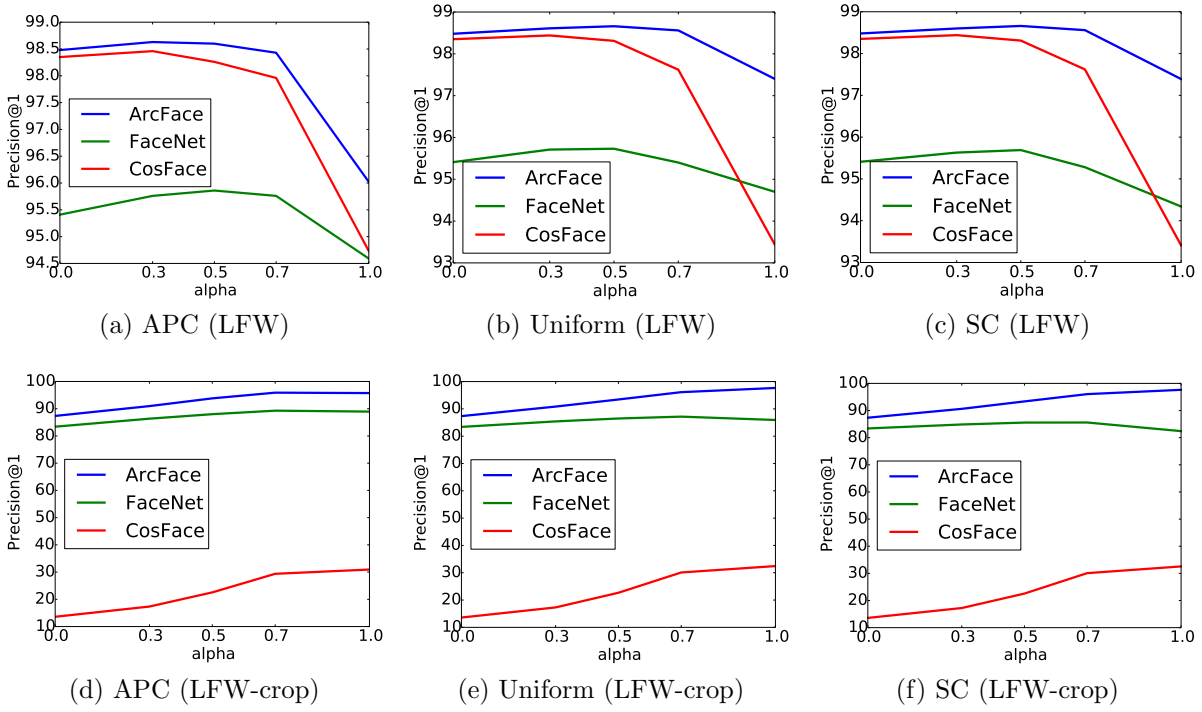
Table 5.2: Comparison of five feature-weighting techniques for ArcFace [22] patch embeddings on LFW [93] and LFW-crop datasets. Performance is often slightly better on a 8×8 grid than on a 4×4. Our 2-stage approach consistently outperforms the vanilla Stage 1 alone and approaches closely the maximum re-ranking precision at $k = 100$.

Experiment We use the grid size of 8×8, i.e. the better setting from the previous ablation study (Sec. 5.2.1). For each pair of images, we linearly combine their patch-level EMD (θ_{EMD}) and the image-level cosine distance (θ_{Cosine}) as:

$$\theta = \alpha \times \theta_{\text{EMD}} + (1 - \alpha) \times \theta_{\text{Cosine}} \tag{5.1}$$

Sweeping across $\alpha \in \{0, 0.3, 0.5, 0.7, 1\}$, we find that changing α has a marginal effect on the P@1 on LFW. That is, the P@1 changes in [95, 98.5] with the lowest accuracy being 95 when EMD is exclusively used, i.e. $\alpha = 1$ (see Fig. 5.1a). In contrast, for LFW-crop, we find the accuracy to monotonically increases as we increase α (Fig. 5.1b). That is, **the**

higher the contribution of patch-wise similarity, the better re-ranking accuracy on the challenging randomly-cropped queries. We choose $\alpha = 0.7$ as the best and default choice for all subsequent FI experiments. Interestingly, our proposed distance (Eq. 5.1) also yields a state-of-the-art face *verification* result on MLFW [86] (Sec. 5.2.2).



(a) APC (LFW) (b) Uniform (LFW) (c) SC (LFW)

(d) APC (LFW-crop) (e) Uniform (LFW-crop) (f) SC (LFW-crop)

Figure 5.1: The P@1 of our 2-stage FI when sweeping across $\alpha \in \{0, 0.3, 0.5, 0.7, 1.0\}$ for linearly combining EMD and cosine distance on LFW (top row; a–c) and LFW-crop images (bottom row; d–f) of all feature weighting (APC, Uniform, and SC).

Patch-wise EMD for ranking or re-ranking

Given that re-ranking using EMD at the patch-embedding space substantially improves the precision of FI compared to Stage 1 alone (Tab. 5.2), here, we test performing such patch-wise EMD sorting at Stage 1 instead of Stage 2.

Experiment That is, we test ranking images using EMD at the patch level instead of the standard cosine distance at the image level. Performing patch-wise EMD at Stage 1 is significantly slower than our 2-stage approach, e.g., **∼12 times slower** (729.20s vs. 60.97s, in total, for 13,233 queries). That is, Sinkhorn is a slow, iterative optimization method and

ArcFace	Method		Time (s)	P@1	RP	MAP@R
(a) LFW	APC	EMD at Stage 1	268.96	83.35	76.97	73.81
		Ours	**60.03**	**98.60**	**78.63**	**78.22**
	SC	EMD at Stage 1	196.50	97.85	77.92	77.29
		Ours	**77.32**	**98.66**	**78.74**	**78.35**
	Uniform	EMD at Stage 1	191.47	97.85	77.91	77.29
		Ours	**77.79**	**98.66**	**78.73**	**78.35**
	LMK	EMD at Stage 1	178.67	98.13	78.18	77.61
		Ours	**77.79**	**98.66**	**78.73**	**78.35**
(b) LFW-crop vs. LFW	APC	EMD at Stage 1	729.20	55.53	44.06	38.57
		Ours	**60.97**	**96.10**	**76.58**	**74.56**
	SC	EMD at Stage 1	266.74	98.57	76.20	74.30
		Ours	**60.39**	96.19	**78.05**	**76.20**
	Uniform	EMD at Stage 1	259.84	**98.62**	76.19	74.28
		Ours	**61.81**	96.26	**78.08**	**76.25**

Table 5.3: Comparison of performing patch-wise EMD ranking at Stage 1 vs. our proposed 2-stage FI approach (i.e. cosine similarity ranking in Stage 1 and patch-wise EMD re-ranking in Stage 2). In both cases, EMD uses 8×8 patches. EMD at Stage 1 is the method of using EMD to rank images directly (instead of the regular cosine similarity) and there is no Stage 2 (re-ranking). For our method, we choose the same setup of $\alpha = 0.7$. Our 2-stage approach does not only outperform using EMD at Stage 1 but is also ∼2-4 × faster. The run time is the total for all **13,214 queries** for both (a) and (b). The result supports our choice of performing EMD in Stage 2 instead of Stage 1.

the EMD at Stage 2 has to sort only $k = 100$ (instead of 13,233) images. In addition, FI by comparing images patch-wise using EMD at Stage 1 yields consistently worse accuracy than our 2-stage method under all feature-weighting techniques (see Tab. 5.3 for details).

5.2.2 Additional Results

To demonstrate the generality and effectiveness of our 2-stage FI, we take the best hyperparameter settings ($\alpha = 0.7$; APC) from the ablation studies (Sec. 5.2.1) and use them for three different models (ArcFace [22], CosFace [89], and FaceNet [70]), which have different grid sizes.

We test the three models on five different OOD query types: (1) faces wearing masks or (2) sunglasses; (3) profile faces; (4) randomly cropped faces; and (5) adversarial faces.

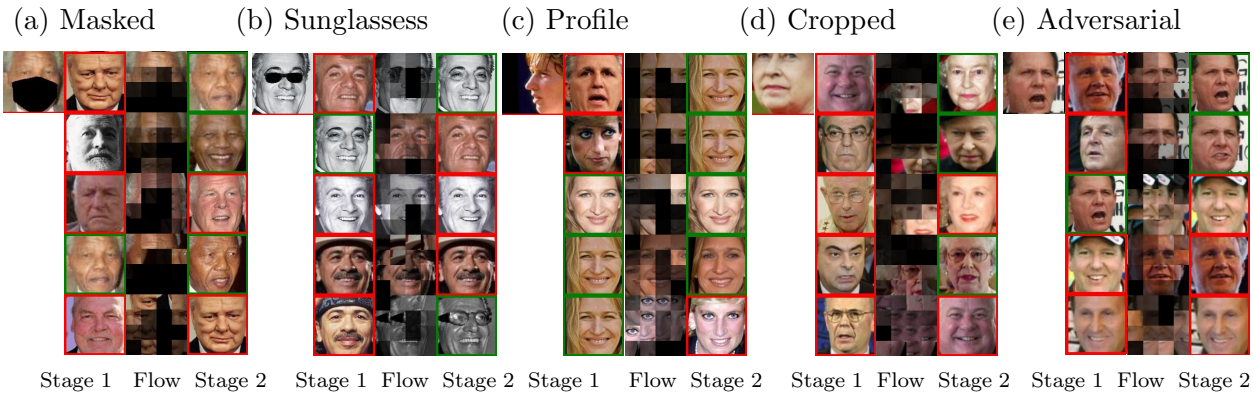


Figure 5.2: Figure in a similar format to that of Fig. 3.1. Our re-ranking based on patch-wise similarity using ArcFace (4×4 grid; APC) pushes more relevant gallery images higher up (here, we show top-5 results), improving face identification precision under various types of occlusions. The "Flow" visualization intuitively shows the patch-wise reconstruction of the query (top-left) given the highest-correspondence patches (i.e. largest flow) from a gallery face. The darker a patch, the lower the flow. For example, despite being masked out ∼50% of the face (a), Nelson Mandela can be correctly retrieved as Stage 2 finds gallery faces with similar forehead patches. See Fig. 3.2 for a similar figure as the results of running our method with an 8×8 grid (i.e. smaller patches), which yields slightly better precision (Tab. 5.2).

Identifying occluded faces

Experiment We perform our 2-stage FI on three datasets: CFP [71], CALFW [102], and AgeDB [55]. 12,173-image CALFW and 16,488-image AgeDB have age-varying images of 4,025 and 568 identities, respectively. CFP has 500 people, each having 14 images (10 frontal and 4 profile).

To test our models on challenging OOD queries, in CFP, we use its 2,000 profile faces in CFP as queries and its 5,000 frontal faces as the gallery. To create OOD queries using CFP[1], CALFW, and AgeDB, we automatically occlude all images with masks and sunglasses by detecting the landmarks of eyes and mouth using dlib and overlaying black sunglasses or a mask on the faces (see examples in Fig. 3.1). We also take these three datasets and create randomly cropped queries (as for LFW-crop in Sec. 5.2.1). For all datasets, we test identifying occluded query faces given the original, unmodified gallery. That is, for every query, there is ≥ 1 matching gallery image.

[1]We only apply masks and sunglasses on the frontal images of CFP.

Dataset	Model	Method	P@1	RP	M@R
CALFW (Mask)	ArcFace	ST1	96.81	53.13	51.70
		Ours	**99.92**	**57.27**	**56.33**
	CosFace	ST1	98.54	43.46	41.20
		Ours	**99.96**	**59.85**	**58.87**
	FaceNet	ST1	77.63	39.74	36.93
		Ours	**96.67**	**45.87**	**44.53**
CALFW (Sunglass)	ArcFace	ST1	51.11	29.38	26.73
		Ours	**54.95**	**30.66**	**27.74**
	CosFace	ST1	45.20	25.93	22.78
		Ours	**49.67**	**26.98**	**24.12**
	FaceNet	ST1	21.68	13.70	10.89
		Ours	**25.07**	**15.04**	**12.16**
CALFW (Crop)	ArcFace	ST1	79.13	43.46	41.20
		Ours	**92.57**	**47.17**	**45.68**
	CosFace	ST1	10.99	6.45	5.43
		Ours	**25.99**	**12.35**	**11.13**
	FaceNet	ST1	79.47	44.40	41.99
		Ours	**85.71**	**45.91**	**43.83**
AgeDB (Mask)	ArcFace	ST1	96.15	39.22	30.41
		Ours	**99.84**	39.22	**33.18**
	CosFace	ST1	98.31	38.17	31.57
		Ours	**99.95**	**39.70**	**33.68**
	FaceNet	ST1	75.99	22.28	14.95
		Ours	**96.53**	**24.25**	**17.49**
AgeDB (Sunglass)	ArcFace	ST1	84.64	51.16	44.99
		Ours	**87.06**	50.40	44.27
	CosFace	ST1	68.93	34.90	27.30
		Ours	**75.97**	**35.54**	**28.12**
	FaceNet	ST1	56.77	27.92	20.00
		Ours	**61.21**	**28.98**	**21.11**
AgeDB (Crop)	ArcFace	ST1	79.92	32.66	26.19
		Ours	**92.92**	**32.93**	**26.60**
	CosFace	ST1	10.11	4.23	2.18
		Ours	**19.58**	**4.95**	**2.76**
	FaceNet	ST1	80.80	31.50	24.27
		Ours	**86.74**	**31.51**	**24.32**

Table 5.4: When the queries (from CALFW [102] and AgeDB [55]) are occluded by masks, sunglasses, or random cropping, our 2-stage method (8×8 grid; APC) is substantially more robust to the Stage 1 alone baseline (ST1) with up to +13% absolute gain (e.g. P@1: 79.13 to 92.57). The conclusions are similar for other feature-weighting methods (see **??** and **??**).

Results First, for all three models and all occlusion types, i.e. due to masks, sunglasses, crop, and self-occlusion (profile queries in CFP), **our method consistently outperforms the traditional Stage 1 alone** approach under all three precision metrics (Tables 5.4, 5.5, & 5.6).

Second, across all three datasets, we find the **largest improvement** that our Stage 2 provides upon the Stage 1 alone is when the queries are **randomly cropped or masked** (Tab. 5.4). In some cases, the Stage 1 alone using cosine distance is not able to retrieve any

Dataset	Model	Method	P@1	RP	M@R
CFP (Profile)	ArcFace	ST1	84.84	71.09	67.35
		Ours	**84.94**	70.31	66.36
	CosFace	ST1	71.64	58.87	54.81
		Ours	71.64	**59.24**	**55.23**
	FaceNet	ST1	75.71	61.78	56.30
		Ours	**76.38**	61.69	56.19

Table 5.5: Our 2-stage approach based on ArcFace features (8×8 grid; APC) performs slightly better than the Stage 1 alone (ST1) baseline at P@1 when the query is a rotated face (i.e. profile faces from CFP [71]). See Tab. 5.6 for the results of occlusions on CFP.

relevant examples among the top-5 but our re-ranking manages to push three relevant faces into the top-5 (Fig. 5.2d).

Third, we observe that for faces with masks or sunglasses, APC interestingly often excludes the mouth or eye regions from the fully-visible gallery faces when computing the EMD patch-wise similarity with the corresponding occluded query (**??**). The same observation can be seen in the visualizations of the most similar patch pairs, i.e. highest flow, for our same 2-stage approach that uses either 4×4 grids (Fig. 5.2 and Fig. 3.1) or 8×8 grids (Fig. 3.2).

Robustness to adversarial images

Adversarial examples pose a huge challenge and a serious security threat to computer vision systems [43, 57] including FI [73, 103]. Recent research suggests that the patch representation may be the key behind ViT impressive robustness to adversarial images [5, 72, 52]. Motivated by these findings, we test our 2-stage FI on TALFW [103] queries given an original 13,233-image LFW gallery.

Experiment TALFW contains 4,069 LFW images perturbed adversarially to cause face verifiers to mislabel [103].

Results Over the entire TALFW query set, we find our re-ranking to consistently outperform the Stage 1 alone under all three metrics (Tab. 5.7). Interestingly, the improvement (of ∼2 to 4 points under P@1 for three models) is larger than when tested on the original LFW queries (around 0.12 in Tab. 5.2a), verifying our patch-based re-ranking robustness when

Dataset	Model	Method	P@1	RP	M@R
CFP (Mask)	ArcFace	Stage 1	96.65	69.88	66.67
		APC	**99.78**	**76.07**	**74.20**
		Uniform	**99.78**	**76.41**	**74.34**
		SC	**99.78**	**76.23**	**74.08**
	CosFace	Stage 1	92.52	66.14	62.73
		APC	**94.22**	**69.56**	**66.66**
		Uniform	**94.38**	**70.34**	**67.59**
		SC	**94.32**	**70.45**	**67.72**
	FaceNet	Stage 1	83.96	54.82	49.01
		APC	**97.48**	**61.58**	**57.35**
		Uniform	**95.63**	**58.71**	**53.96**
		SC	**93.09**	**57.30**	**52.15**
CFP (Sunglass)	ArcFace	Stage 1	91.54	70.63	67.21
		Uniform	**93.10**	**71.75**	**68.33**
		APC	**94.06**	**71.05**	**67.89**
		SC	**92.92**	**71.69**	**68.24**
	CosFace	Stage 1	88.72	65.93	61.97
		APC	**82.22**	**60.33**	**54.25**
		Uniform	**85.28**	**61.89**	**56.65**
		SC	**86.04**	**62.53**	**57.45**
	FaceNet	Stage 1	69.02	50.58	43.26
		APC	**74.98**	**52.98**	**46.14**
		Uniform	**69.18**	**51.46**	**43.87**
		SC	**67.90**	**50.67**	**43.02**
CFP (Crop)	ArcFace	Stage 1	91.34	65.13	61.37
		Uniform	**98.16**	**70.77**	**67.80**
		APC	**97.96**	**67.51**	**64.15**
		SC	**98.04**	**70.78**	**67.78**
	CosFace	Stage 1	17.06	10.51	8.02
		SC	**34.60**	**15.69**	**12.96**
		Uniform	**34.50**	**15.63**	**12.90**
		APC	**32.22**	**15.07**	**12.23**
	FaceNet	Stage 1	95.20	72.70	69.43
		APC	**97.34**	**72.63**	**69.47**
		Uniform	**96.54**	**72.78**	**69.56**
		SC	**96.02**	**72.22**	**68.88**
CFP (Profile)	ArcFace	Stage 1	84.84	71.09	67.35
		Uniform	**86.13**	**72.19**	**68.58**
		APC	**85.56**	**71.60**	**67.84**
		SC	**86.18**	**72.22**	**68.59**
	CosFace	Stage 1	71.64	58.87	54.81
		SC	**71.74**	**59.27**	**55.27**
		Uniform	**71.74**	**59.21**	**55.22**
		APC	**71.64**	**59.24**	**55.23**
	FaceNet	Stage 1	75.71	61.78	56.30
		APC	**76.38**	61.69	56.19
		Uniform	**76.33**	61.47	55.89
		SC	**76.22**	61.35	55.74

Table 5.6: More results of our 2-stage approach based on ArcFace features (8×8 grid), CosFace features (6× 7), and FaceNet features (3 × 3) across all feature weighting methods which perform slightly better than the Stage 1 alone (ST1) baseline at P@1 when the query is a rotated face (i.e. profile faces from CFP [71]).

Dataset	Model	Method	P@1	RP	M@R
TALFW [103] vs. LFW [93]	ArcFace	ST1	93.49	81.04	80.35
		Ours	**96.64**	**82.72**	**82.10**
	CosFace	ST1	96.49	83.57	82.99
		Ours	**99.07**	**85.48**	**85.03**
	FaceNet	ST1	95.33	79.24	78.19
		Ours	**97.26**	**80.33**	**79.39**

Table 5.7: Our re-ranking (8×8 grid; APC) consistently improves the precision over Stage 1 alone (ST1) when identifying adversarial TALFW [103] images given an in-distribution LFW [93] gallery. The conclusions also carry over to other feature-weighting methods.

queries are perturbed with very small noise. That is, our approach can improve FI precision when the perturbation size is either small (adversarial) or large (e.g. masks).

Re-ranking rivals finetuning on masked images

While our approach does not involve re-training, a common technique for improving FI robustness to occlusion is data augmentation, i.e. re-train the models on occluded data in addition to the original data. Here, we compare our method with data augmentation on masked images.

Finetuning hyperparameters We describe here the hyperparameters used for finetuning ArcFace on our CASIA dataset augmented with masked images (see Fig. 5.3 for some samples).

- Training on $907,459$ facial images (masks and non-masks).
- Number of epochs is 12.
- Optimizer: SGD.
- Weight decay: $5e^{-4}$
- Learning rate: 0.001
- Margin: $m = 0.5$

- Feature scale: $s = 30.0$

See details in the published code base: code

Experiment To generate augmented, masked images, we follow [6] to overlay various types of masks on CASIA images to generate ∼415K masked images. We add these images to the original CASIA training set, resulting in a total of ∼907K images (10,575 identities). We finetune ArcFace on this dataset with the same original hyperparameters [2]. We train three models and report the mean and standard deviation (Tab. 5.8).

For a fair comparison, we evaluate the finetuned models and our no-training approach on the MLFW dataset [86], instead of our self-created masked datasets. That is, the query set has 11,959 MLFW masked-face images and the gallery is the entire 13,233-image LFW.

Results First, we find that finetuning ArcFace improves its accuracy in FI under Stage 1 alone (Tab. 5.8; 39.79 vs. 41.64). Yet, our 2-stage approach still substantially outperforms Stage 1 alone, both when using the original and the finetuned ArcFace (Tab. 5.8; 48.23 vs. 41.64). Interestingly, we also test using the finetuned model in our DeepFace-EMD framework and finds it to approach closely the best no-training result (46.21 vs. 48.23).

ArcFace	Method	P@1	RP	M@R
Pre-trained	(a) ST1	39.79	35.10	33.32
	(b) Ours	**48.23**	**41.43**	**39.71**
Finetuned	(c) ST1	41.64± 0.16	34.67± 0.24	32.66± 0.25
	(d) Ours	46.21 ± 0.27	38.65 ± 0.26	36.73± 0.26

Table 5.8: Our 2-stage approach (b) using ArcFace (8×8 grid; APC) substantially outperforms Stage 1 alone (a) on identifying masked images of MLFW given the unmasked gallery of LFW. Interestingly, our method (b) also outperforms Stage 1 alone when ArcFace has been finetuned on masked images (c). In (c), we report the mean and std over three finetuned models.

Face Verification on MLFW

In the main text, we find that DeepFace-EMD is effective in face *identification* given many types of OOD images. Here, we also evaluate DeepFace-EMD for face *verification* of MLFW [86], a recent benchmark that consists of masked LFW faces. As in common

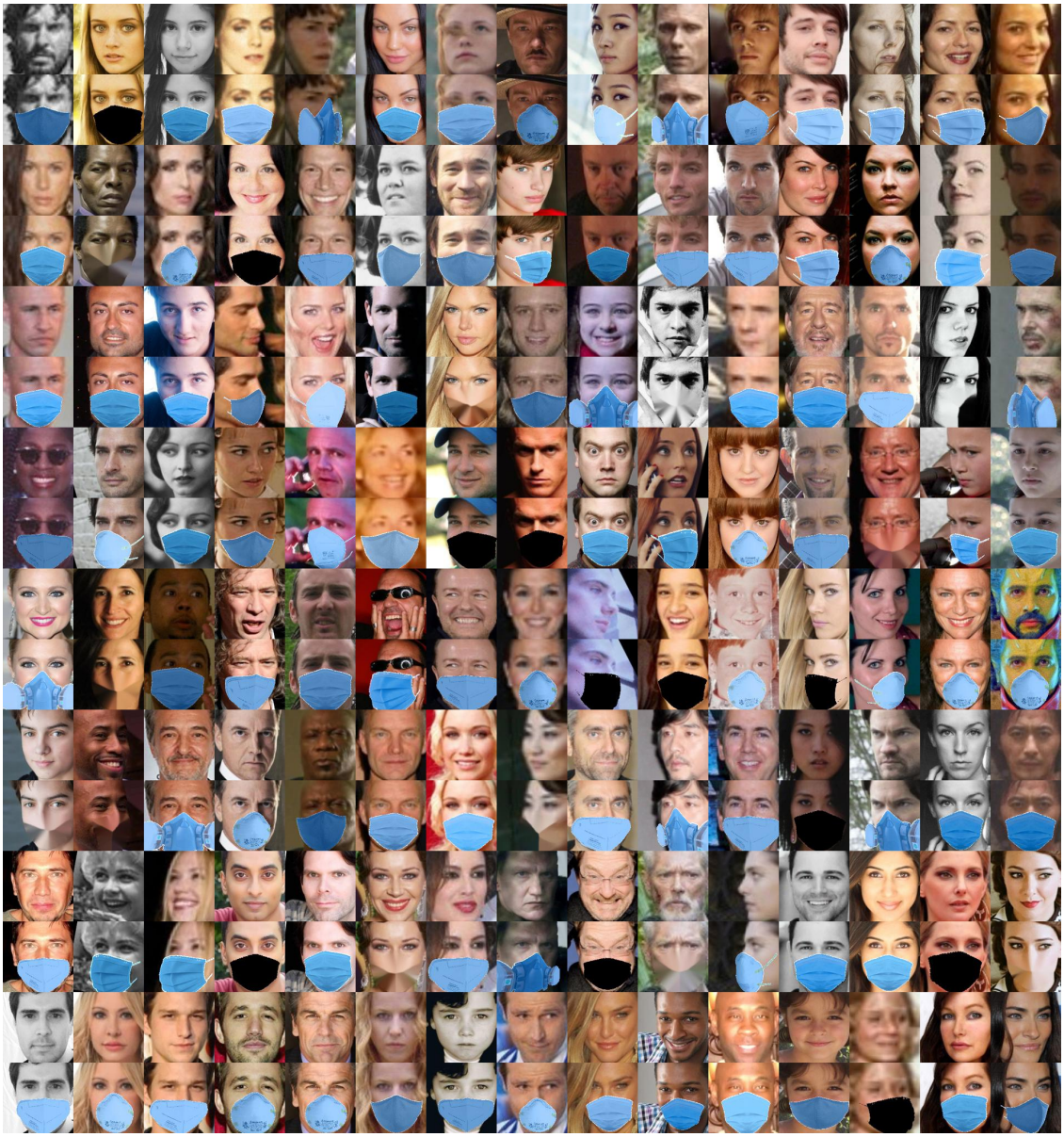


Figure 5.3: Our CASIA dataset augmented with masked images (generated following the method by [6]) for fine-tuning ArcFace.

verification setups of LFW [70, 50, 86], given pairs of face images and their similarity scores predicted by a verification system, we find the optimal threshold that yields the best accuracy. Here, we follow the setup in [86] to enable a fair comparison. First of all, we reproduce Table 3 in [86], which evaluate face verification accuracy on 6,000 pair of MLFW images. Then, we run our DeepFace-EMD distance function (Eq. 5.1). We found that using our proposed distance consistently improves on face *verification* for all three PyTorch models in [86].

Interestingly, with DeepFace-EMD, **we obtained a state-of-the-art result** (91.17%) on MLFW (see Tab. 5.9).

Models in MLFW Table 3 [76]	Method	MLFW
Private-Asia, R50, ArcFace	[76]	74.85%
	+ DeepFaceEMD	**76.50%**
CASIA, R50, CosFace	[76]	82.87%
	+ DeepFaceEMD	**87.17%**
MS1MV2, R100, Curricularface	[76]	90.60%
	+ DeepFaceEMD	**91.17%**

Table 5.9: Using our proposed similarity function consistently improves the face verification results on MLFW (i.e. OOD masked images) for models reported in Wang et al. [86]. We use pre-trained models and code by [86].

5.3 Face-ViT: Fast and Interpretable Face Recognition for Out-Of-Distribution Data Using Vision Transformers (ViTs)

5.3.1 Ablation Studies

For model understanding and parameter selection, we conduct two major ablation studies for networks with different settings: (1) Cross-attention 2-image vs. no-cross-attention 1-image, for both in-distribution data and OOD (Sec. 5.3.1), and (2) With cross-attention, 2-output linear vs. 1-output *[CLS]* (Sec. 5.3.1). In addition, we provide a study for how to select the depth and the head of Transformers (Sec. 5.3.1).

Datasets. We run face verification experiments on two datasets: the in-distribution LFW [93] and the masked-face-occlusion MLFW [86]. The face verification task has 6,000 pairs (3000 positives and 3000 negatives, a total of 12,000 images). For the hybrid models (C, and D), we used the pre-trained ResNet18 ArcFace model [22]. Images are aligned and cropped to 128×128 by the MTCNN algorithm [70]. Inputs are normalized to $[0, 1]$ by subtracting 127.5 and dividing by 127.5. For Model V, images are cropped to 112×112 with original RGB values in $[0, 255]$. All models are trained on a clean and processed CASIA Webface database [93].

(a) LFW (b) TALFW (c) MLFW

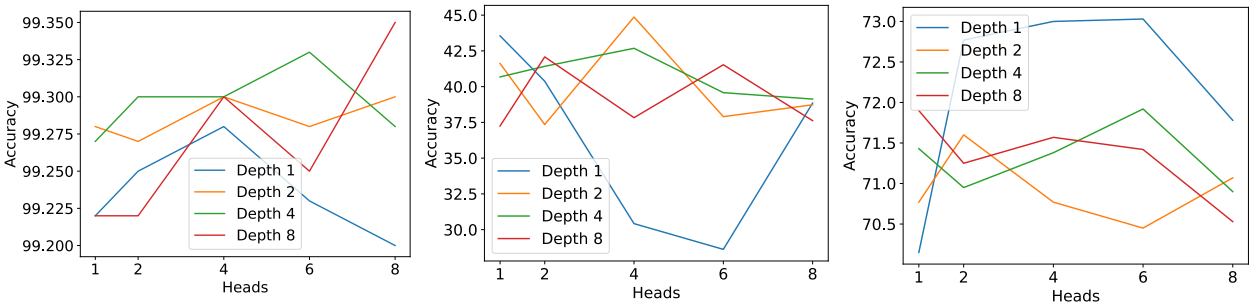


Figure 5.4: The efficiency of settings of depths and heads for the network (**H2L**) within different domains. For LFW, the depth of 1 achieved comparable accuracy with a depth of 8 (e.g. very small difference of 0.075 %). In TALFW, with depths of 1 and 2 and heads of 1 and 4 respectively, the accuracy outperforms the accuracy of depths of 4 and 8. For face masks in MLFW, the depth of 1 consistently outperforms the other settings. Therefore, using a low depth of **1** or **2** for contextual information design can gain good performance.

Model training. We train models with a batch size of 320 images and a learning rate of $1e^{-6}$ for the first warm-up epoch and $1e^{-5}$ in the remaining 49 epochs. For Transformer settings, the models are trained with depth $= 1, 2, 4, 8$ and head $= 1, 2, 4, 6, 8$. For CNN backbones in hybrid-ViTs, we do not update the parameters. For ArcFace loss [22], hyper-parameters are as follow.

- Margin: $m = 0.5$
- Feature scale: $s = 30.0$

All experiments are run on eight 40GB A100 SXM GPUs.

The low depth's efficiency

For the 2-image/output hybrid-ViT (H2L), adding the Transformer layer at the top of CNN can improve the performance. However, increasing the number of depths and heads can lead to redundant computation in the models while showing no meager improvements in terms of accuracy. Here, we evaluate the effects of different values of depths and heads to select the potential settings in face problems.

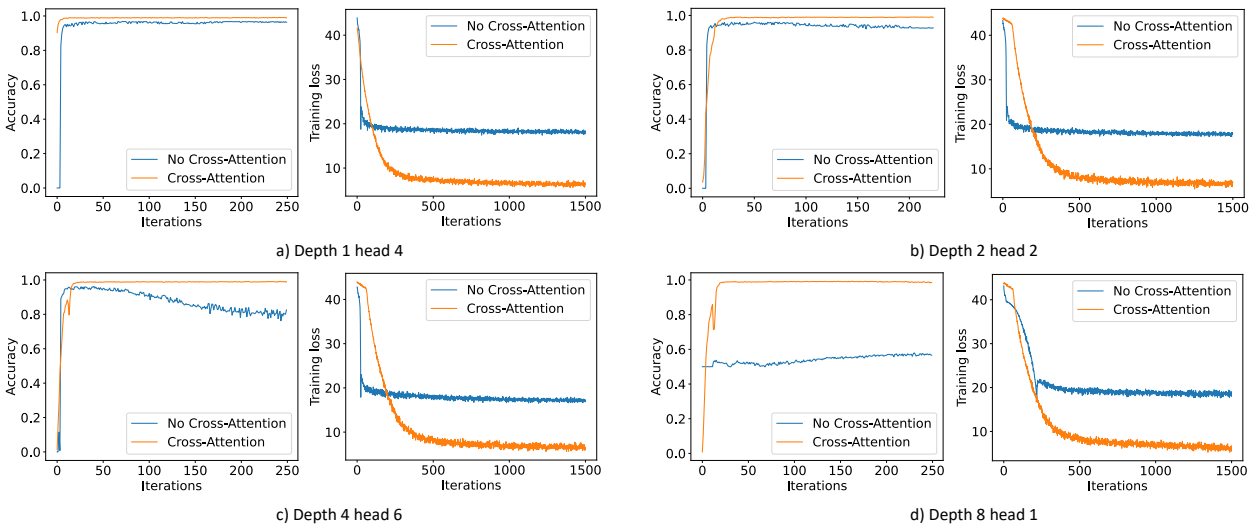


Figure 5.5: Comparison in accuracy and convergence between training **H1** (No-cross-attention) vs. **H2L** (Cross-attention) architectures on LFW [93]. For different network settings, 2-input-image achieves better accuracy and more stable training when leveraging patch-wise cross-image attention.

Experiment As mentioned above, we use depths $= 1, 2, 4, 8$ and heads $= 1, 2, 4, 6, 8$. We report the accuracy of H2L for face verification on LFW, TALFW, and MLFW datasets.

Results First, we observe that on LFW, the low depth of 1 achieves lower performance. However, it still **outperforms the CNN model** (99.28% Fig. 5.4 (head=4) vs 98.02% in Fig. 5.5). Moreover, a lower value of depth and head **achieves comparable results** compared to higher values (e.g. 99.22% with depth of 1, head of 1, vs. 99.34% with depth of 8, head of 8). Second, for the TALFW dataset, hybrid-ViTs (H2L) also **achieves comparable accuracy** (with d=1, h=1), i.e. performing well with low depth and head values for face adversarial. Third, for the MLFW dataset, the depth of 1 **outperforms** the other higher-depth value models.

The cross-attention 2-image ViT outperforms the 1-image

To investigate our hypothesis that using cross-attention can improve the performance in face recognition, we compare our proposed 2-image (cross-attention) model with the 1-image (no-cross-attention) one.

model	depth	head	LFW	MLFW
C CNN	-	-	98.02	70.75
V ViT	20	8	97.77	57.62
H1 Hybrid-ViT (1-image)	1	4	96.38	56.00
	2	2	96.13	57.85
	4	6	96.20	57.75
	8	1	58.00	57.92
H2L Hybrid-ViT (2-image)	1	4	**99.28**	**73.00**
	2	2	**99.27**	**71.60**
	4	6	**99.30**	**71.92**
	8	1	**99.22**	**71.90**

Table 5.10: Comparison of 1-image (no-cross-attention) and 2-image (cross-attention). 2-image hybrid model H2L outperforms 1-image models (C, V, and H1) on in-distribution (LFW) and occlusion OOD (MLFW) domains. In addition, the accuracy of the low depth is similar to higher depth so that we can use the low depths. Therefore, we can rule out models: C, V, and H1, and choose the lower depth of H2L.

2-image Hybrid-ViT	depth	head	LFW	MLFW
H2 CLS (1-output)	1	1	90.45	48.40
	1	2	96.38	53.55
	1	4	97.47	56.88
	2	1	92.47	52.52
H2L 2-Linear (2-output)	1	1	**99.22**	**70.15**
	1	2	**99.25**	**72.77**
	1	4	**99.28**	**73.00**
	2	1	**99.28**	**70.77**

Table 5.11: Model H2L with 2-output features outperforms H2 (CLS output) on both LFW and MLFW.

Experiment. For Model V, we use a depth of 20 and a head of 8. For Model V & H1, we use *[CLS]* outputs to extract 512-dimension features. For Model H2L, we use the remaining 2-output with 512-dimension embeddings. All features are learned with the ArcFace loss function [22] to classify identities.

Results. First, we find that the 2-image (cross-attention) model outperforms the 1-image (no-cross-attention) one significantly on the LFW and MLFW datasets, showing that cross-image information is useful for handling OOD data (Tab. 5.10). For example, in LFW, the accuracy of H2L (depth=4, head=6) increases $\sim 3.14\%$ (model H1), $\sim 1.5\%$ (Model V), and $\sim 1.25\%$ (CNN). Furthermore, the 2-image model H2L substantially provides more useful

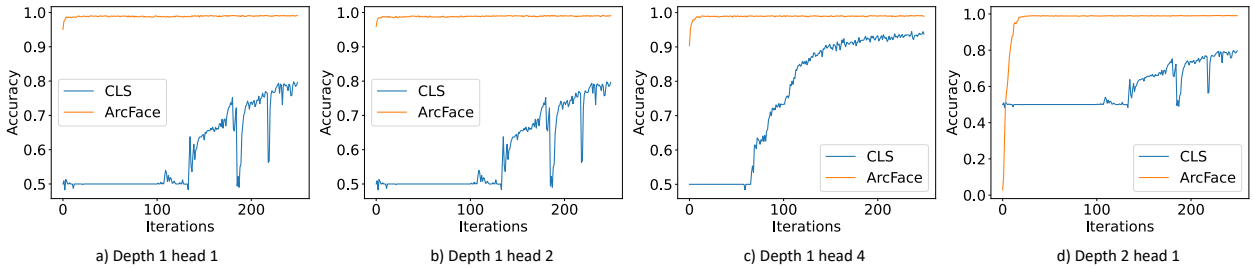


Figure 5.6: Training performance of CLS (model **H2**) and ArcFace hybrid-ViT (model **H2L**) on LFW. Model **H2L** consistently outperforms and achieves more stability in the training process.

similarity information than the 1-image model for OOD distribution on MLFW (Tab. 5.10; Model H2L - 73% vs. C-70.75%, H1-57.92%, and V-57.62%).

Second, interestingly, we find that the hybrid models (H1 & H2L) can achieve higher precision with a depth of only 1, i.e. adding an efficient shallow layer to Transformers can improve performance (e.g. on LFW, 99.28% H2L vs. 98.02 % of H1). We deduce the same statement when comparing it with the ViT model (V). In contrast, the 1-image no-cross-attention model has worse performance with the in-distribution LFW (see Fig. 5.5) and the OOD MLFW (Tab. 5.10). With a higher depth of 8, model H1 becomes worse in LFW (Tab. 5.10 H2L-99.22% vs. H1-58.00%)

Cross-Attention: The 2-linear-output ViT outperforms the 1-output [CLS]

The previous Transformer-based FI works [27, 24] usually use an extra learnable embedding *[CLS]*, discarding the remaining embeddings that may contain helpful cross-image information. Here, we experiment with the 1-output *[CLS]* (model H2) and 2-output (model H2L) to study how the embeddings can improve performance.

Experiment. In the 1-output *[CLS]*, we deploy binary cross entropy loss to classify identities. We train Transformers with depths of 1 and 2.

Results. First, we find that the 2-linear-output model H2L consistently outperforms the 1-output *[CLS]* model H2 on LFW and MLFW (Tab. 5.11), verifying that the remaining embeddings cross-image information between two images are helpful to improve models. In

LFW (in-distribution), the 2-output model improves the accuracy by +8.55 points (Tab. 5.11; from 90.45% of H2 to 99.22% of H2L). In the out-of-distribution masked-face image (MLFW) datasets, the improvement is even more significant when the accuracy increases by +21.75 points (Tab. 5.11; 48.40% of H2 vs. 70.15% of H2L).

Second, the training of the 2-output Model H2L performs better and is more stable than the 1-output Model H2 in only a few iterations (Fig. 5.6). For instance, the 1-output *[CLS]* Model H2 only achieves 80% in accuracy over LFW while the 2-output model H2L can reach 99% in accuracy within fewer iterations (Fig. 5.6a, b, and d).

To sum up, we can improve model performance on OOD by using a low depth of 1, which saves computational costs and proves that H2L performs better in both in-distribution and OOD domains. In addition, with higher depths, H2 performs worse.

5.3.2 Main Results

In Sec. 5.3.2, we experiment on different OOD query types including masks, sunglasses, and adversarial faces. Here, we select the best settings from ablation studies in **??** including depth of 1 and head of 1, 2, or 6. In Sec. 5.3.2, we show that our model has a faster time complexity compared with other layer types. Sec. 5.3.3 discusses our model's face explainability. To boost the performance, our proposed Model H2L can be used in a 2-stage fashion like DeepFaceEMD, i.e. selecting the top 100 Stage 1's candidates ($k = 100$) with CNN w.r.t cosine similarity scores and then re-ranking these candidates with cross-image features — 2 outputs from Transformers. We also re-use a combination of two stages with $\alpha = 0.7$, which works best for occlusion cases [63]. The models trained with settings mentioned in Sec. 5.3.1 are reported with 2 stages (ST1 and ST2) compared with the original ArcFace and DeepFaceEMD. The results are computed by three metrics: P@1, RP, and M@R [101, 56]. For the details of these metrics, see Sec. 5.1.

Comparable accuracy

Experiment. We demonstrate our models for FI on two datasets: CALFW [102] and AgeDB [55]. The 12,173 CALFW images and 16,488 AgeDB images have age-varying of 4,025 and 568 identities, respectively. We re-use OOD queries of these datasets from DeepFaceEMD [63] consisting of masks and sunglasses.

Results. First, in ST1, 2-image (model H2L) achieves comparable accuracy with the original ArcFace [22]. In the AgeDB dataset, ST1's P@1 of model H2L improves around +2 points over model C on Mask (98.73% vs. 96.31%; Tab. 5.12) and Sunglasses (86.01% vs. 84.64%; Tab. 5.12), increasing the accuracy on occlusion in the cross-age domain.

Second, ST2 of Model H2L significantly outperforms ST1 (e.g. CALFW (mask) **99.29**% vs. **95.58**% P@1 in Tab. 5.12) and achieves better results compared with DeepFaceEMD in sunglass images (ST2 on RP and M@R metrics in Tab. 5.12), verifying the boost performance in the 2-stage process.

Comparable robustness

Experiment. To illustrate the effectiveness of adversarial attacks, we run the experiment on the TALFW dataset [103]. TALFW contains 13,233 images perturbed adversarially to fool face models.

Results. First, in ST2, model H2L achieves better results than model H1 on all 3 metrics, P@1 (H2L-94.03% vs. H1-93.49%), RP (H2L-81.63% vs. H1-81.04%), and M@R (H2L-81.09% vs. H1-80.35%). See the last row of Tab. 5.12), verifying that our proposed model H2L also improves the precision in adversarial images with a re-ranking algorithm. Second, DeepFace-EMD (model D) achieves the best results in all metrics both ST1 and ST2 (see the last row of Tab. 5.12). These results show that these models (models H2L & D) are robust to adversarial images, which is a grand challenge in computer vision [43, 57].

dataset	name	model	stage	depth	head	P@1	RP	M@R
CALFW (Mask)	C	CNN	ST1	-	-	95.58	51.59	50.01
	H2L	Hybrid-ViT	ST1	1	2	95.03	43.70	42.36
	D	DeepFaceEMD	ST2	-	-	**99.79**	**56.77**	**55.75**
	H2L	Hybrid-ViT	ST2	1	2	99.29	51.00	50.01
CALFW (Sunglasses)	C	CNN	ST1	-	-	51.11	29.38	26.73
	H2L	Hybrid-ViT	ST1	1	6	50.23	28.08	25.15
	D	DeepFaceEMD	ST2	-	-	**54.95**	30.66	27.74
	H2L	Hybrid-ViT (ST2)	ST2	1	6	54.00	**31.00**	**27.87**
AgeDB (Mask)	C	CNN	ST1	-	-	96.31	39.22	30.41
	H2L	Hybrid-ViT	ST1	1	1	98.73	20.68	14.86
	D	DeepFaceEMD	ST2	-	-	**99.84**	**39.22**	**33.18**
	H2L	Hybrid-ViT	ST2	1	1	99.28	33.93	26.69
AgeDB (Sunglasses)	C	CNN	ST1	-	-	84.64	51.16	45.00
	H2L	Hybrid-ViT	ST1	1	2	86.01	49.34	43.03
	D	DeepFaceEMD	ST2	-	-	**87.06**	50.04	44.27
	H2L	Hybrid-ViT	ST2	1	2	86.75	**51.16**	**44.88**
TALFW vs. LFW	C	CNN	ST1	-	-	93.49	81.04	80.35
	H2L	Hybrid-ViT	ST1	1	2	94.59	77.66	77.00
	D	DeepFaceEMD	ST2	-	-	**96.64**	**82.72**	**82.10**
	H2L	Hybrid-ViT	ST2	1	2	94.03	81.63	81.09

Table 5.12: Face occlusions and adversarial images. Model H2L achieves comparable accuracy on the OOD of CALFW and AgeDB compared to CNN and DeepFace-EMD [63].

Layer type	Complexity per layer	Actual runtime (s)	Maximum path Length
C. Convolutional	$O(k \cdot n \cdot d^2)$	-	$O(\log_k n)$
V. ViT, Self-Attention	$O(n^2 \cdot d)$	-	$O(1)$
V. Self-Attention (restricted)	$O(r \cdot n \cdot d^2)$	-	$O(n/r)$
H2L Hybrid-ViT	$O(k \cdot n \cdot d^2 + n^2 \cdot d)$	**24.33**	$O(\log_k n)$
D. DeepFace-EMD [63]	$O(k \cdot n \cdot d^2 + n^3 \cdot \log n)$ [75]	53.35	$O(1)$

Table 5.13: Time complexity of different type layers. n is the number of patches, d is the dimension of embeddings, k is the kernel size of convolutions, and r is the size of the neighborhood in restricted self-attention.

Faster inference time

We evaluate the time complexity of different layers. For vanilla ViT and CNN, the time complexity is mentioned in the Transformer network [84] (see Tab. 5.13). Hybrid-ViTs consist of convolutional neural networks and low-depth self-attention at the top. Hence,

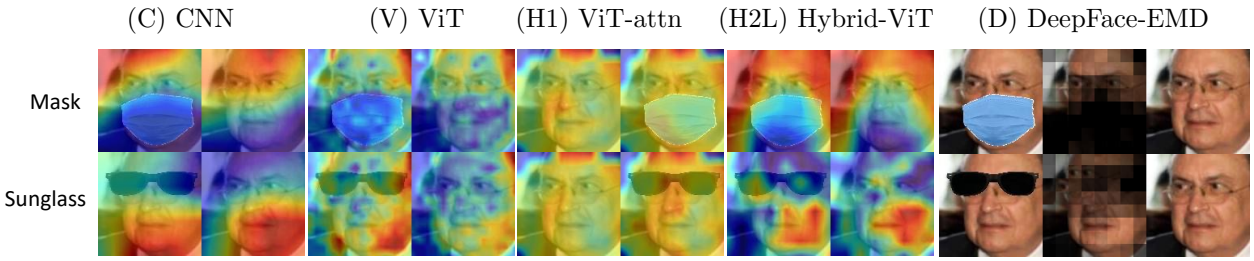


Figure 5.7: Comparison of face models' explainability on **LFW** OOD domains. ViT-attn is visualized through the method of Chefer et al. [16]. Our proposed H2L can highlight the important area in images (e.g. eyes, mouth, etc.) and remove occluded areas (e.g. mask and sunglasses). In contrast, Model V contains noisy heatmaps and H1 does not provide any interpretable clues of how two faces match.

time complexity can be added by convolutional layers and self-attention layers (the last row in Tab. 5.13). The run-time complexities of Model C, V, H2L, and D are shown in Tab. 5.13. Our Model H2L has a lower complexity, $O(n^2)$, than that of DeepFace-EMD, $O(n^3)$. In practice, Model H2L performs at least 2 times faster than Model D when used as the re-ranking process (ST2) in face identification (see Tab. 5.14 and Fig. 5.9). Moreover, in ST2, DeepFace-EMD is slow to solve EMD for higher dimension patch-wise similarity [63] while hybrid-ViT simply computes the cosine similarity of cross-image features and low-depth Transformers, i.e. enhancing the scalability. For example, in AgeDB (sunglasses), the computation is sped up to $\sim$ **3**$\times$ for 16,409 sunglass-query images in settings of 8×8 patches (see Tab. 5.14 for details). Therefore, model H2L is a good choice for more scalable architectures.

Dataset	Model	# of queries	Time (seconds)	Depth	Head	P@1	RP	M@R
CALFW (Mask)	D DeepFaceEMD [63]	11,914	53.35	-	-	**99.79**	**56.77**	**55.75**
	H2L Hybrid-ViT		**24.33**	1	2	99.29	51.00	50.01
CALFW (Sunglass)	D DeepFaceEMD [63]	12,173	73.90	-	-	**54.95**	30.66	27.74
	H2L Hybrid-ViT		**29.10**	1	6	54.00	**31.00**	**27.87**
AgeDB (Mask)	D DeepFaceEMD [63]	15,629	72.42	-	-	**99.84**	**39.22**	**33.18**
	H2L Hybrid-ViT		**34.44**	1	1	99.28	33.93	26.69
AgeDB (Sunglass)	D DeepFaceEMD [63]	16,409	90.40	-	-	**87.06**	50.04	44.27
	H2L Hybrid-ViT		**33.01**	1	2	86.75	**51.16**	**44.88**

Table 5.14: Actual running times and performance for ST2 computation in face identification under occlusion. Compared to DeepFace-EMD (D), the computation of hybrid-ViTs (H2L) is significantly faster. For example, for 11,914 query images of the CALFW (mask), H2L runs at least **2** times faster.

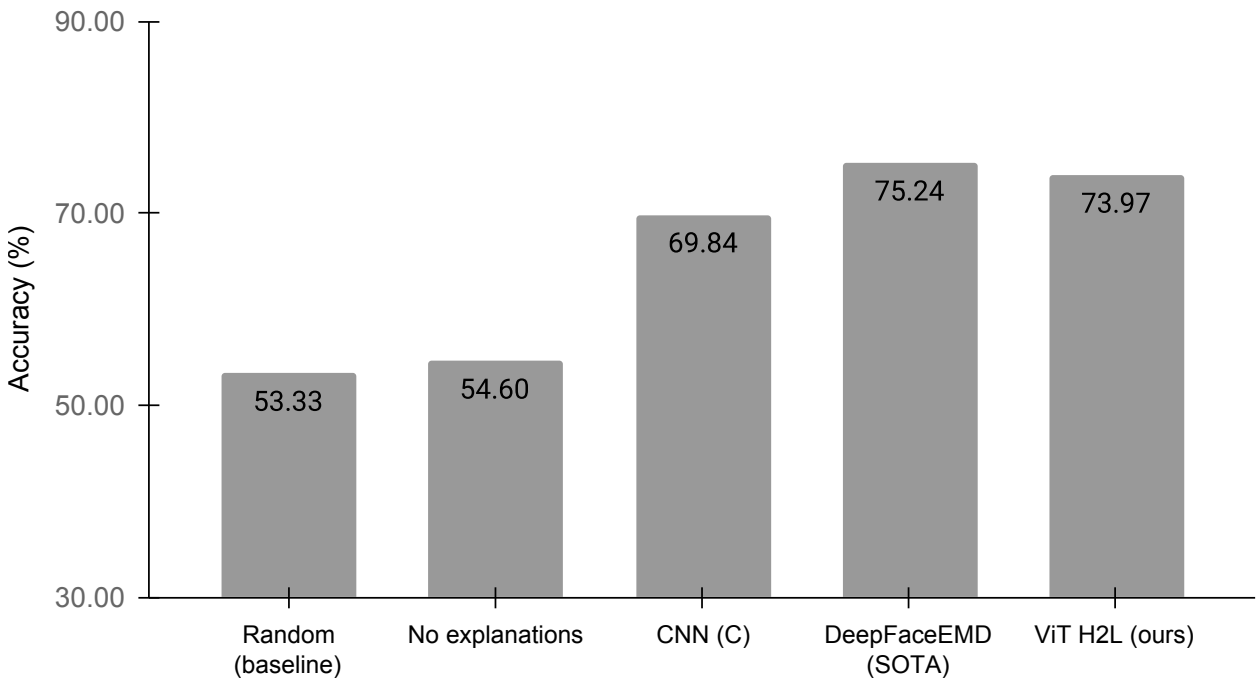


Figure 5.8: Human explainability across various networks. The mean and standard deviation of the accuracy of 21 users when presented with 4 explanations: Cross-correlation (CC) method on CNNs [77]; flow visualization in DeepFace-EMD [63]; CC on 2-image Hybrid-ViT; and a baseline of no explanations. The explanations of Model D and H2L result in substantially higher user accuracy than those of Model C and the No-explanation baseline, which is close to the random baseline of 53.33%.

5.3.3 Better model explanation by human evaluation

As face identification systems in the real world are often customer-facing [67, 37, 33, 12, 69], we study how CNNs (model C), 1-image ViTs (model V), 2-image Hybrid-ViT (model H2L), and DeepFace-EMD (model D) help users in understanding face verification results. For each image pair, we generate a visual explanation from a model (examples in Fig. 5.7), and ask a user to look at both images and the explanation and decide whether the two faces are of the same person.

Experiment. Similar to [101, 63], we use the cross correlation method from [77] to generate similarity heatmaps for the CNNs and ViTs. This method produces a heatmap by taking

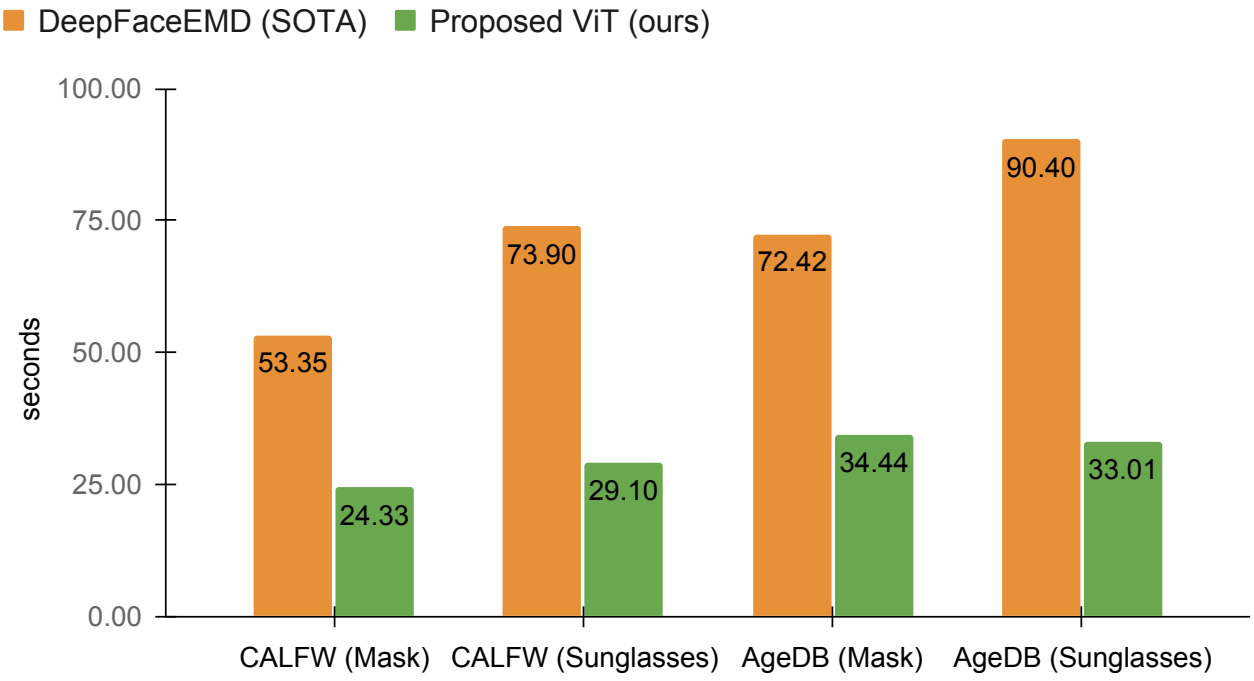


Figure 5.9: Actual running times for the re-ranking computation in face identification under occlusion. Our proposed model is at least two times faster than the state-of-the-art DeepFace-EMD [63] over all the datasets.

the dot product between every patch embedding of image 1 and the global average pooling feature of image 2. For DeepFace-EMD, we plot their flow visualizations as in [63].

The explanation heatmaps are generated for models C, H2L, and D using their last convolutional layers, which have the same spatial dimension of 8×8. For model V, the spatial dimension of the heatmap is 14×14. In preliminary experiments, we find the raw cross-attention matrices at the first layer of the ViT model uninformative to users (see Fig. 5.7; ViT-attn). Therefore, we use cross-correlation (CC) [77] to generate explanations for ViTs (Fig. 5.7; ViT).

We recruit 21 participants who are graduate students across multiple institutions in the U.S., Vietnam, and China. For each user, we provide them 5 training examples and 15 pairs of images per method (i.e. 15 pairs × 4 methods = 60 pairs in total). We randomly mask

and place a pair of sunglasses on each image. Sec. 5.3.3 presents specific examples and how we design for user study.

Results. First, we find that users without any model explanations score an average accuracy of 54.60%, i.e. near random chance (53.33%). This suggests that the face verification task is challenging to users (which is consistent with the qualitative feedback obtained from users).

Second, all model explanations are useful in improving user accuracy. Model H2L and D are most useful to users who score 73.97% and 75.24% respectively. Interestingly, these explanations of Model H2L and D, which leverage cross-image interaction, are more useful than the CC explanations of CNNs, which do not allow cross-image interaction (69.84% user accuracy; Fig. 5.8). In sum, consistent with the accuracy-based analysis in Sec. 5.3.2 & Sec. 5.3.2, our user study finds models with cross-image interaction (Model H2L and F) have higher explainability to users.

User study samples

Fig. 5.10, Fig. 5.12, Fig. 5.12, and Fig. 5.13 are specific examples for our design for user study. These figures are only the first pages to instruct users for each approach.

Here, we experiment with 4 approaches: no explanation, Hybrid-ViT, CNNs [77], and EMD [63].

Welcome to our user study for face matching!! User study evaluates on domains of masked, sunglass , and normal faces. To make Yes/No decisions, please look at both faces AND the middle visualizations, which highlight the key similarities between the two faces. Try your best to verify matching facial pairs. To answer the question, you can highlight your answers: Yes/No. Use 1st page as the examples. Start in the 2nd page.

Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

Figure 5.10: User study for no-explanation method.

Welcome to our user study for face matching!! User study evaluates on domains of masked, sunglass
, and normal faces. To make Yes/No decisions, please look at both faces AND the middle visualizations, which highlight the key similarities between the two faces. Try your best to verify matching facial pairs. To answer the question, you can highlight your answers: Yes/No. Use 1st page as the examples. Start in the 2nd page.

Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

Figure 5.11: User study for Hybrid-ViT method.

Welcome to our user study for face matching!! User study evaluates on domains of masked, sunglass
, and normal faces. To make Yes/No decisions, please look at both faces AND the middle visualizations, which highlight the key similarities between the two faces. Try your best to verify matching facial pairs. To answer the question, you can highlight your answers: Yes/No. Use 1st page as the examples. Start in the 2nd page.

Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

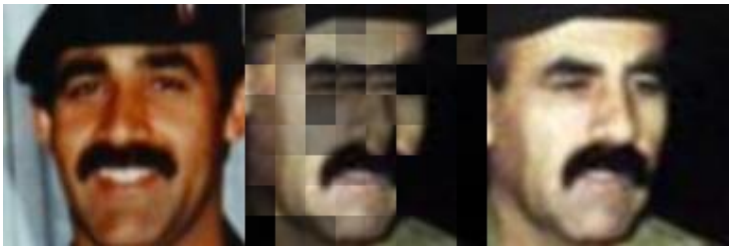
Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No

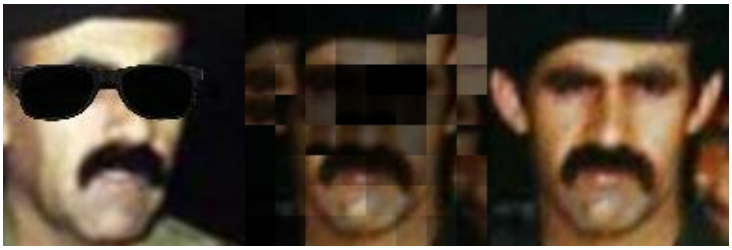
Figure 5.12: User study for CNNs method.

Welcome to our user study for face matching!! User study evaluates on domains of masked, sunglass
, and normal faces. To make Yes/No decisions, please look at both faces AND the middle visualizations, which highlight the key similarities between the two faces. Try your best to verify matching facial pairs. To answer the question, you can highlight your answers: Yes/No. Use 1st page as the examples. Start in the 2nd page.



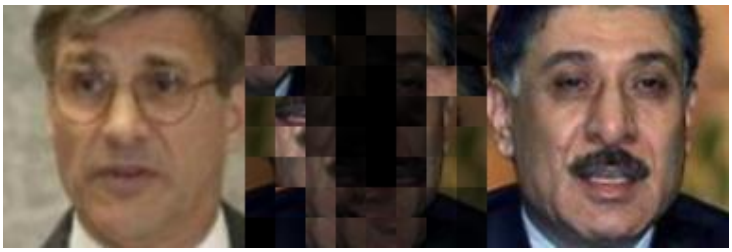
Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No



Are these two faces of the same person? Your answer: Yes / No

Are these two faces of the same person? Your answer: Yes / No



Are these two faces of the same person? Your answer: Yes / No

Figure 5.13: User study for the EMD method.

Chapter 6

Conclusion & Future Works

We proposed two novel methods: DeepFace-EMD and Face-ViT based on structure similarity for interpretable face recognition systems.

For **DeepFace-EMD**, solving patch-wise EMD via Sinkhorn is slow, which may prohibit it from being used to sort a much larger image sets (see run-time reports in Tab. 5.3). Furthermore, here, we used EMD on two distributions of equal weights; however, the algorithm can be used for unequal-weight cases [66, 19], which may be beneficial for handling occlusions. While substantially improving FI accuracy under the four occlusion types (i.e., masks, sunglasses, random crops, and adversarial images), re-ranking is only marginally better than Stage 1 alone on ID and profile faces, which is interesting to understand deeper in future research.

Instead of using pre-trained models, it might be interesting to re-train new models explicitly on patch-wise correspondence tasks, which may yield better patch embeddings for our re-ranking. In sum, we propose DeepFace-EMD, a 2-stage approach for comparing images hierarchically: First at the image level and then at the patch level. DeepFace-EMD shows impressive robustness to occluded and adversarial faces and can be easily integrated into existing FI systems in the wild.

For **Face-ViT** First, we find that using models that leverage cross-image interaction as the re-ranker substantially improves FI accuracy under occlusion and adversarially perturbed queries. Second, we train a 2-image Hybrid-ViT model that not only achieves similar accuracy but also two times faster than state-of-the-art models. Note that the 1-image models, which do not use cross-image interaction, are still the fastest to run in practice since they enable caching of the image embeddings. Finally, we find visual explanations of cross-image

interaction models are substantially more useful in improving lay-user face verification accuracy than not having explanations. We also perform the first study in the literature comparing state-of-the-art FI approaches in three main criteria: accuracy, complexity, and explainability.

Significance. Face identification in the wild is essentially a hard, ill-posed zero-shot image retrieval task. We hope our work can inspire more explorations in the use of ViTs and EMD for face identification and to improve the speed of this system in the real world.

Future work. The performance of hybrid-ViTs is still slightly lower than that of DeepFace-EMD. It would be possible to tune ViT hyperparameters [8] for higher accuracy and incorporate sparsity into the attention mechanism of ViT for improved inference speed.

References

[1] Nijeer parks was arrested due to a false facial recognition match - cnn. https://www.cnn.com/2021/04/29/tech/nijeer-parks-facial-recognition-police-arrest/index.html. (Accessed on 07/09/2021).

[2] ronghuaiyang/arcface-pytorch. https://github.com/ronghuaiyang/arcface-pytorch. (Accessed on 11/16/2021).

[3] S. Adam and H. Kashmir. Barred from grocery stores by facial recognition. https://www.seattletimes.com/business/barred-from-grocery-stores-by-facial-recognition/. (Accessed on 26/08/2023).

[4] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[5] Anonymous. Patches are all you need? In *Submitted to The Tenth International Conference on Learning Representations*, 2022. under review.

[6] A. Anwar and A. Raychowdhury. Masked face recognition for secure authentication. *ArXiv*, abs/2008.11104, 2020.

[7] H. Bao, L. Dong, and F. Wei. BEiT: BERT pre-training of image transformers. *ICLR 2022*, 2022.

[8] L. Beyer, X. Zhai, and A. Kolesnikov. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022.

[9] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4000–4009, 2017.

[10] S. Bharadwaj, M. Vatsa, and R. Singh. Aiding face recognition with social context association rule based re-ranking. In *IEEE International Joint Conference on Biometrics*, pages 1–8. IEEE, 2014.

[11] S. Black, A. Stylianou, R. Pless, and R. Souvenir. Visualizing paired image similarity in transformer networks. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1534–1543, 2022.

[12] N. M. Burke. Michigan man wrongfully accused with facial recognition urges congress to act. https://www.detroitnews.com/story/news/politics/2021/07/13/house-panel-hear-michigan-man-wrongfully-accused-facial-recognition/7948908002/. (Accessed on 11/09/2021).

[13] J. Cai, H. Han, J. Cui, J. Chen, L. Liu, and S. K. Zhou. Semi-supervised natural face de-occlusion. *IEEE Transactions on Information Forensics and Security*, 16:1044–1057, 2020.

[14] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74, 2018.

[15] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2018.

[16] H. Chefer, S. Gur, and L. Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.

[17] H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

[18] C.-F. R. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 357–366, 2021.

[19] S. Cohen. *Finding color and shape patterns in images*. stanford university, 1999.

[20] J. Cui, F. Wen, and X. Tang. Real time google and live image search re-ranking. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 729–732, 2008.

[21] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013.

[22] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[25] J. Dong, L. Zhang, H. Zhang, and W. Liu. Occlusion-aware gan for face de-occlusion in the wild. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[26] X. Dong, S. Kim, Z. Jin, J. Y. Hwang, S. Cho, and A. Teoh. Open-set face identification with index-of-max hashing by learning. *Pattern Recognition*, 103:107277, 2020.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[28] E. Efendić, P. P. Van de Calseyde, and A. M. Evans. Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, 157:103–114, 2020.

[29] A. Gailey. Facial recognition tech at hartsfield-jackson wins over most international delta customers - atlanta business chronicle. `https://www.bizjournals.com/atlanta/news/2019/06/25/facial-recognition-tech-at-hartsfield-jackson-wins.html`. (Accessed on 11/09/2021).

[30] C. Gartenberg. Apple's face id with a mask works so well, it might end password purgatory. `https://www.theverge.com/2022/2/2/22912677/apple-face-id-mask-update-ios-15-4-beta-hands-on-impressions`. (Accessed on 26/08/2023).

[31] S. Ge, C. Li, S. Zhao, and D. Zeng. Occluded face recognition in the wild by identity-diversity inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3387–3397, 2020.

[32] J. Guo, X. Zhu, Z. Lei, and S. Z. Li. Face synthesis for eyeglass-robust face recognition. In *Chinese Conference on biometric recognition*, pages 275–284. Springer, 2018.

[33] D. Harwell. Wrongfully arrested man sues detroit police following false facial-recognition match - the washington post. `https://www.washingtonpost.com/technology/2021/04/13/facial-recognition-false-arrest-lawsuit/`. (Accessed on 11/09/2021).

[34] D. Harwell and T. Craig. The fbi's capitol riot investigation used surveillance technology that advocates say threatens civil liberties - the washington post. `https://www.washingtonpost.com/technology/2021/04/02/capitol-siege-arrests-technology-fbi-privacy/`. (Accessed on 04/11/2021).

[35] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[36] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong. A regularized correntropy framework for robust pattern recognition. *Neural computation*, 23(8):2074–2100, 2011.

[37] K. Hill. Flawed facial recognition leads to arrest and jail for new jersey man - the new york times. `https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html`. (Accessed on 11/09/2021).

[38] W. Kim, B. Son, and I. Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[39] W. Kim, B. Son, and I. Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 18–24 Jul 2021.

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

[41] R. Koner, P. Sinhamahapatra, K. Roscher, S. Günnemann, and V. Tresp. Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976*, 2021.

[42] S. Kumar, S. Chakrabarti, and S. Roy. Earth mover's distance pooling over siamese lstms for automatic short answer grading. In *IJCAI*, pages 2046–2052, 2017.

[43] A. Kurakin, I. Goodfellow, S. Bengio, et al. Adversarial examples in the physical world, 2016.

[44] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

[45] E. Levina and P. Bickel. The earth mover's distance is the mallows distance: Some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 251–256. IEEE, 2001.

[46] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

[47] X.-X. Li, D.-Q. Dai, X.-F. Zhang, and C.-X. Ren. Structured sparse error coding for face recognition with occlusion. *IEEE transactions on image processing*, 22(5):1889–1900, 2013.

[48] Z. Li, W. Liu, D. Lin, and X. Tang. Nonparametric subspace analysis for face recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 961–966. IEEE, 2005.

[49] S. Liao and L. Shao. Transmatcher: Deep image matching through transformers for generalizable person re-identification. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[51] N. Lupu, L. Selios, and Z. Warner. A new measure of congruence: The earth mover's distance. *Political Analysis*, 25(1):95–113, 2017.

[52] K. Mahmood, R. Mahmood, and M. Van Dijk. On the robustness of vision transformers to adversarial examples. *arXiv preprint arXiv:2104.02610*, 2021.

[53] S. Mia. The pandemic is testing the limits of face recognition. `https://www.technologyreview.com/2021/09/28/1036279/pandemic-unemployment-government-face-recognition/`. (Accessed on 26/08/2023).

[54] R. Min, A. Hadid, and J.-L. Dugelay. Improving the recognition of faces occluded by facial accessories. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 442–447. IEEE, 2011.

[55] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, page 5, 2017.

[56] K. Musgrave, S. Belongie, and S.-N. Lim. A metric learning reality check. In *ECCV*, pages 681–699. Springer, 2020.

[57] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[58] H. J. Oh, K. M. Lee, and S. U. Lee. Occlusion invariant face recognition using selective local non-negative matrix factorization basis images. *Image and Vision computing*, 26(11):1515–1523, 2008.

[59] E. Osherov and M. Lindenbaum. Increasing cnn robustness to occlusions by reducing filter support. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 550–561, 2017.

[60] C. Peng, N. Wang, J. Li, and X. Gao. Re-ranking high-dimensional deep local representation for nir-vis face recognition. *IEEE Transactions on Image Processing*, 28(9):4553–4565, 2019.

[61] Y. Peng, C. Fang, and X. Chen. Using earth mover's distance for audio clip retrieval. In *Pacific-Rim Conference on Multimedia*, pages 405–413. Springer, 2006.

[62] H. Phan, C. Le, V. Le, Y. He, and A. Nguyen. Fast and interpretable face recognition for out-of-distribution data using vision transformers (vits). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[63] H. Phan and A. Nguyen. Deepface-emd: Re-ranking using patch-wise earth mover's distance improves out-of-distribution face identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[64] P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[65] H. Qiu, D. Gong, Z. Li, W. Liu, and D. Tao. End2end occluded face recognition by masking corrupted features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[66] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[67] T. Ryan-Mosley. The new lawsuit that shows facial recognition is officially a civil rights issue — mit technology review. `https://www.technologyreview.com/2021/04/14/1022676/robert-williams-facial-recognition-lawsuit-aclu-detroit-police/`. (Accessed on 04/15/2021).

[68] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018.

[69] M. Sato. The pandemic is testing the limits of face recognition — mit technology review. `https://www.technologyreview.com/2021/09/28/1036279/pandemic-unemployment-government-face-recognition/`. (Accessed on 11/09/2021).

[70] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[71] S. Sengupta, J. Cheng, C. Castillo, V. Patel, R. Chellappa, and D. Jacobs. Frontal to profile face verification in the wild. *IEEE Conference on Applications of Computer Vision*, February 2016.

[72] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh. On the adversarial robustness of visual transformers. *arXiv preprint arXiv:2103.15670*, 2021.

[73] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016.

[74] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3013–3020. IEEE, 2012.

[75] S. Shirdhonkar and D. W. Jacobs. Approximate earth mover's distance in linear time. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[76] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 773–782, 2019.

[77] A. Stylianou, R. Souvenir, and R. Pless. Visualizing deep similarity networks. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 2029–2037. IEEE, 2019.

[78] A. Stylianou, R. Souvenir, and R. Pless. Visualizing deep similarity networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[79] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015.

[80] T. Swearingen and A. Ross. Lookalike disambiguation: Improving face identification performance at top ranks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10508–10515. IEEE, 2021.

[81] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[82] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. Training data-efficient image transformers and distillation through attention. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021.

[83] D. S. Trigueros, L. Meng, and M. Hartnett. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. *Image and Vision Computing*, 79:99–108, 2018.

[84] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[85] W. Wan and J. Chen. Occlusion robust face recognition based on mask learning. In *2017 IEEE international conference on image processing (ICIP)*, pages 3795–3799. IEEE, 2017.

[86] C. Wang, H. Fang, Y. Zhong, and W. Deng. Mlfw: A database for face recognition on masked faces. *arXiv preprint arXiv:2109.05804*, 2021.

[87] F. Wang and L. J. Guibas. Supervised earth mover's distance learning and its computer vision applications. In *European Conference on Computer Vision*, pages 442–455. Springer, 2012.

[88] H. Wang, J. Shen, Y. Liu, Y. Gao, and E. Gavves. Nformer: Robust person re-identification with neighbor transformer. *arXiv preprint arXiv:2204.09331*, 2022.

[89] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018.

[90] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2008.

[91] X. Xu, N. Sarafianos, and I. A. Kakadiaris. On improving the generalization of face recognition in the presence of occlusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 798–799, 2020.

[92] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In *CVPR 2011*, pages 625–632. IEEE, 2011.

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

[94] C. Zhang, Y. Cai, G. Lin, and C. Shen. Deepemd: Differentiable earth mover's distance for few-shot learning, 2020.

[95] C. Zhang, Y. Cai, G. Lin, and C. Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[96] C. Zhang, Y. Cai, G. Lin, and C. Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12203–12213, 2020.

[97] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[98] X. Zhang, M. Jiang, Z. Zheng, X. Tan, E. Ding, and Y. Yang. Understanding image retrieval re-ranking: A graph neural network perspective. *arXiv preprint arXiv:2012.07620*, 2020.

[99] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan. Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE Transactions on Image Processing*, 27(2):778–790, 2017.

[100] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan. Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE Transactions on Image Processing*, 27:778–790, 2018.

[101] W. Zhao, Y. Rao, Z. Wang, J. Lu, and J. Zhou. Towards interpretable deep metric learning with structural matching. In *ICCV*, 2021.

[102] T. Zheng, W. Deng, and J. Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017.

[103] Y. Zhong and W. Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2020.

[104] Y. Zhong and W. Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021.

[105] S. Zhou, J. Luo, J. Zhou, and X. Ji. Asarcface: Asymmetric additive angular margin loss for fairface recognition. In *ECCV Workshops*, 2020.

[106] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma. Face recognition with contiguous occlusion using markov random fields. In *2009 IEEE 12th international conference on computer vision*, pages 1050–1057. IEEE, 2009.

[107] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.