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

Hai Phan\textsuperscript{1,2}  
pthai1204@gmail.com  
\textsuperscript{1}Auburn University  
\textsuperscript{2}Carnegie Mellon University  

Anh Nguyen\textsuperscript{1}  
anh.ng8@gmail.com

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. Here, we propose a re-ranking approach that compares two faces using the Earth Mover’s Distance on the deep, spatial features of image patches. 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.

1. Introduction

Who shoplifted from the Shinola luxury store in Detroit [4]? Who are you to receive unemployment benefits [5] or board an airplane [1]? Face identification (FI) today is behind the answers to such life-critical questions. Yet, the technology can make errors, leading to severe consequences, e.g., people wrongly denied of unemployment benefits [5] or falsely arrested [2–4,7]. 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. First, a deep feature extractor may not have seen a normal, non-celebrity person before during 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 [44, 54] (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. For example, face verification accuracy may notoriously drop significantly (from 99.38% to 81.12% on LFW) given an occluded query face (Fig. 1b–d) or adversarial queries [73]. Our main findings are:

- Interestingly, the OOD accuracy can be substantially improved via a 2-stage approach (see Fig. 2): 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) [45] (Sec. 3 & Sec. 4).
- 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. 4). 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 [59], our re-ranking method (no training) rivals the ArcFace models finetuned directly on masked images (Sec. 4.3).

To our knowledge, our work is the first to demonstrate the remarkable effectiveness of EMD for comparing OOD, occluded and adversarial images at the deep feature level.

2. Methods

2.1. Problem formulation

To demonstrate the generality of our method, we adopt the following simple FI formulation as in [19, 33, 71]: Identify the person in a query image by ranking all 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. 2). Here, our image embeddings are extracted from the last linear layer of all three models and are all \( \in \mathbb{R}^{512} \).

Stage-1: Ranking A common 1-stage face identification [33, 47, 61] 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. 2). 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 compare faces in two hierarchical stages (Fig. 2), first at a coarse, image level and then a fine-grained, patch level.

Via an ablation study (Sec. 3), 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).

2.2. Networks

Pre-trained models We use three state-of-the-art PyTorch models of ArcFace, FaceNet, and CosFace pre-trained on CASIA [65], VGGFace2 [14], and CASIA, respectively. Their architectures are ResNet-18 [24], Inception-ResNet-v1 [56], and 20-layer SphereFace [33], respectively. See Sec. S1 for more details on network architectures and implementation in PyTorch.

Image pre-processing For all networks, we align and crop input images following the 3D facial alignment in [11] (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. 1) are pre-processed. Using MTCNN, the default pre-processing of all three networks, does not change the results substantially (Sec. S5).

2.3. 2-stage hierarchical face identification

To our knowledge, our work is the first to demonstrate the remarkable effectiveness of EMD for comparing OOD, occluded and adversarial images at the deep feature level.

2.4. Earth Mover’s Distance (EMD)

EMD is an edit distance between two set of weighted objects or distributions [45]. Its effectiveness was first demonstrated in measuring pair-wise image similarity based on color histograms and texture frequencies [45] for image retrieval. Yet, EMD is also an effective distance between two text documents [29], probability distributions (where EMD is equivalent to Wasserstein, i.e. Mallows distance) [30], and distributions in many other domains [27, 34, 43]. Here, we propose to harness EMD as a distance between two faces, i.e. two sets of weighted facial features.

Let \( Q = \{ (q_1, w_{q_1}), \ldots, (q_N, w_{q_N}) \} \) 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 \( Q \) and the set of weighted features of a gallery face \( G = \{ (g_1, w_{g_1}), \ldots, (g_N, w_{g_N}) \} \) is any matrix \( F = (f_{ij}) \in \mathbb{R}^{N \times N} \). 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
Figure 2. 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.

distance between \( (q_i, g_j) \) and \( D = (d_{ij}) \in \mathbb{R}^{N \times N} \) be the ground distance matrix of all pair-wise distances.

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

\[
\text{COST}(Q, G, F) = \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} f_{ij} \tag{1}
\]

s.t. \( f_{ij} \geq 0 \)

\[
\sum_{j=1}^{N} f_{ij} \leq w_{q_i}, \quad \text{and} \quad \sum_{i=1}^{N} f_{ij} \leq w_{g_j}, \quad i, j \in [1, N] \tag{2}
\]

\[
\sum_{i=1}^{N} \sum_{j=1}^{N} f_{ij} = \min \left( \sum_{j=1}^{N} w_{g_j}, \sum_{i=1}^{N} w_{q_i} \right) \tag{3}
\]

As in \([66, 71]\), we normalize the weights of a face such that the total weights of features \( i \text{ is } i.e. \sum_{i=1}^{N} w_{q_i} = \sum_{j=1}^{N} w_{g_j} = 1 \), which is also the total flow in Eq. (4). Note that EMD is a metric iff two distributions have an equal total weight and the ground distance function is a metric \([16]\).

We use the iterative Sinkhorn algorithm \([18]\) to efficiently solve the linear programming problem in Eq. (1), which yields the final EMD between two faces \( Q \) and \( G \).

**Facial features** In image retrieval using EMD, a set of features \( \{ q_i \} \) can be a collection of dominant colors \([45]\), spatial frequencies \([45]\), or a histogram-like descriptor based on the local patches of reference identities \([60]\). Inspired by \([60]\), 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. 1a). Therefore, without facial occlusion, every image patch is supposed to contain useful identity information, which is in contrast to natural photos \([66]\).

Our grid sizes \( H \times W \) for ArcFace, FaceNet, and CosFace are respectively, \( 8 \times 8 \), \( 3 \times 3 \), and \( 6 \times 7 \), which are the corresponding spatial dimensions of their last convolutional layers (see definitions of these layers in Sec. S1). That is, each feature \( q_i \) is an embedding of size \( 1 \times 1 \times C \) where \( C \) is the number of channels (\textit{i.e.} 512, 1792, and 512 for ArcFace, FaceNet, and CosFace, respectively).

**Ground distance** Like \([66, 71]\), we use cosine distance as the ground distance \( d_{ij} \) between the embeddings \( (q_i, g_j) \) of two patches:

\[
d_{ij} = 1 - \frac{\langle q_i, g_j \rangle}{\|q_i\| \|g_j\|} \tag{5}
\]

where \( \langle ., . \rangle \) is the dot product between two feature vectors.

### 2.5. 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. \([66]\) found that it is beneficial to assign lower weight to less informative regions (\textit{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. 1):

\[
w_{q_i} = w_{q_j} = \frac{1}{N}, \text{ where } 1 \leq k \leq N
\]

\[(6)\]

**Average Pooling Correlation (APC)** Instead of uniformly weighting all patch embeddings, an alternative from [66] 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 \( \{g_j\}_N \) of the gallery image:

\[
w_{q_i} = \max (0, \frac{\sum_j N g_j}{N}), w_{q_j} = \max (0, \frac{\sum_i N g_i}{N})
\]

\[(7)\]

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. 3b; blue shades around both the mask and the non-occluded mouth).

**Cross Correlation (CC)** APC [66] is different from CC introduced in [71], 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 [53], is to take the sum of the cosine similarity between the query patch \( q_i \) and every patch in each gallery image \( \{g_j\}_N \) :

\[
w_{q_i} = \max (0, \sum_j N \frac{q_i \cdot g_j}{\|q_i\| \|g_j\|}), w_{g_j} = \max (0, \sum_i N \frac{q_i \cdot g_j}{\|q_i\| \|g_j\|})
\]

\[(8)\]

We observe that SC often assigns a higher weight to occluded regions e.g., masks and sunglasses (Fig. 3b).

**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 [26] to predict 68 keypoints in each face image (see Fig. 3c) 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. 3c).

![Figure 3. The results of assigning weights to 4×4 patches for ArcFace under three different techniques. Based on per-patch density of detected landmarks (---), LMK (c) often assigns higher weight to the center of a face (regardless of occlusions). In contrast, SC and APC assign higher weight to patches of higher patch-wise and patch-vs-image similarity, respectively. APC tends to down-weight a facial feature (e.g. blue regions around sunglasses or mouth) if its corresponding feature is occluded in the other image (b). In contrast, SC is insensitive to occlusions (a). See Fig. S1 & Fig. S2 for more heatmap examples of feature weighting.](image-url)

### 3. 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. 3.1)? (2) re-ranking using both EMD and cosine distance (Sec. 3.2); and (3) comparing patches or images in Stage 1 (Sec. 3.3).

**Experiment** For all three experiments, we use ArcFace to perform FI on both LFW [65] and LFW-crop. For LFW, we take all 1,680 people who have ≥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. 5d). 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. 1a; Max prec.).

### 3.1. 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. S5), which may impair the patch-wise similarity. Therefore, for each weight-
ing 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. 1a) 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 [66, 71]. This is in stark contrast to the finding in [71] 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. 1b; 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. 1b; 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. 1 vs. Fig. S5).

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. 1a; 98.48 vs. 99.88) while there is a large ~12% headroom in LFW-crop (Tab. 1a; 97.35 vs. 98.71). Interestingly, our re-ranking results approach the upperbound re-ranking precision (e.g. Tab. 1b; 96.26 of Uniform vs. 98.71 Max prec. at k = 100).

### 3.2. 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). 

**Experiment** We use the grid size of 8×8, *i.e.* the better setting from the previous ablation study (Sec. 3.1). For each pair of images, we linearly combine their patch-level EMD (θ_{EMD}) and the image-level cosine distance (θ_{Cosine}) as:

\[
\theta = \alpha \times \theta_{EMD} + (1 - \alpha) \times \theta_{Cosine}
\]  

Sweeping across \(\alpha \in \{0, 0.3, 0.5, 0.7, 1\}\), we find that changing \(\alpha\) 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. 4a). In contrast, for LFW-crop, we find the accuracy to monotonically increases as we increase \(\alpha\) (Fig. 4b). 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. 9) also yields a state-of-the-art face verification result on MLFW [59] (Sec. S4).

### 3.3. 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. 1), here, we test performing such patch-wise EMD sorting at Stage 1 instead of Stage 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>P@1</th>
<th>RP</th>
<th>M@R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1 alone [19]</td>
<td>98.48</td>
<td>78.69</td>
<td>78.29</td>
</tr>
<tr>
<td>Max prec. at k = 100</td>
<td>99.88</td>
<td>81.32</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1. Comparison of five feature-weighting techniques for ArcFace [19] patch embeddings on LFW [65] 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 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 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. S1 for details).

4. 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. 3) and use them for three different models (ArcFace [19], CosFace [61], and FaceNet [47]), 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.

4.1. Identifying occluded faces

Experiment We perform our 2-stage FI on three datasets: CFP [48], CALFW [72], and AgeDB [37]. 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, 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. 1). We also take these three datasets and create randomly cropped queries (as for LFW-crop in Sec. 3). For all datasets, we test identifying occluded query faces given the original, unmodified gallery. That is, for every query, there is $\geq 1$ matching gallery image.

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 2, S8, & S4).

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. 2). In some cases, the Stage 1 alone using cosine distance is not able to retrieve any relevant examples among the top-5 but our re-ranking manages to push three relevant faces into the top-5 (Fig. 5d).

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 (Fig. 3). 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 and Fig. 1) or 8×8 grids (Fig. S5).

4.2. Robustness to adversarial images

Adversarial examples pose a huge challenge and a serious security threat to computer vision systems [28, 39] including FI [50, 73]. Recent research suggests that the

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Method</th>
<th>P@1</th>
<th>RP</th>
<th>M@R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALFW (Mask)</td>
<td>ArcFace</td>
<td>STI</td>
<td>96.81</td>
<td>53.13</td>
<td>51.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>99.92</td>
<td>57.27</td>
<td>56.33</td>
</tr>
<tr>
<td></td>
<td>CosFace</td>
<td>STI</td>
<td>98.54</td>
<td>43.46</td>
<td>41.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>99.96</td>
<td>59.85</td>
<td>58.87</td>
</tr>
<tr>
<td></td>
<td>FaceNet</td>
<td>STI</td>
<td>77.63</td>
<td>39.74</td>
<td>36.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>96.67</td>
<td>45.87</td>
<td>44.53</td>
</tr>
<tr>
<td>CALFW (Sunglass)</td>
<td>ArcFace</td>
<td>STI</td>
<td>51.11</td>
<td>29.38</td>
<td>26.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>54.95</td>
<td>30.66</td>
<td>27.74</td>
</tr>
<tr>
<td></td>
<td>CosFace</td>
<td>STI</td>
<td>45.20</td>
<td>25.93</td>
<td>22.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>49.67</td>
<td>26.98</td>
<td>24.12</td>
</tr>
<tr>
<td></td>
<td>FaceNet</td>
<td>STI</td>
<td>21.68</td>
<td>13.70</td>
<td>10.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>25.07</td>
<td>15.04</td>
<td>12.16</td>
</tr>
<tr>
<td></td>
<td>ArcFace</td>
<td>STI</td>
<td>79.13</td>
<td>43.46</td>
<td>41.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>92.57</td>
<td>47.17</td>
<td>45.68</td>
</tr>
<tr>
<td></td>
<td>CosFace</td>
<td>STI</td>
<td>10.99</td>
<td>6.45</td>
<td>5.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>25.99</td>
<td>12.35</td>
<td>11.13</td>
</tr>
<tr>
<td></td>
<td>FaceNet</td>
<td>STI</td>
<td>79.47</td>
<td>44.40</td>
<td>41.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>85.71</td>
<td>45.91</td>
<td>43.83</td>
</tr>
<tr>
<td>AgeDB (Mask)</td>
<td>ArcFace</td>
<td>STI</td>
<td>96.15</td>
<td>39.22</td>
<td>30.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>99.84</td>
<td>39.22</td>
<td>33.18</td>
</tr>
<tr>
<td></td>
<td>CosFace</td>
<td>STI</td>
<td>98.31</td>
<td>38.17</td>
<td>31.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>99.95</td>
<td>39.70</td>
<td>33.68</td>
</tr>
<tr>
<td></td>
<td>FaceNet</td>
<td>STI</td>
<td>75.99</td>
<td>22.28</td>
<td>14.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>96.53</td>
<td>24.25</td>
<td>17.49</td>
</tr>
<tr>
<td></td>
<td>ArcFace</td>
<td>STI</td>
<td>84.64</td>
<td>51.16</td>
<td>44.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>87.06</td>
<td>50.40</td>
<td>44.27</td>
</tr>
<tr>
<td></td>
<td>CosFace</td>
<td>STI</td>
<td>68.93</td>
<td>34.90</td>
<td>27.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>75.97</td>
<td>35.54</td>
<td>28.12</td>
</tr>
<tr>
<td></td>
<td>FaceNet</td>
<td>STI</td>
<td>56.77</td>
<td>27.92</td>
<td>20.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>61.21</td>
<td>28.98</td>
<td>21.11</td>
</tr>
<tr>
<td>AgeDB (Sunglass)</td>
<td>ArcFace</td>
<td>STI</td>
<td>79.92</td>
<td>32.66</td>
<td>26.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>92.92</td>
<td>32.93</td>
<td>26.60</td>
</tr>
<tr>
<td></td>
<td>CosFace</td>
<td>STI</td>
<td>10.11</td>
<td>4.23</td>
<td>2.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>19.58</td>
<td>4.95</td>
<td>2.76</td>
</tr>
<tr>
<td></td>
<td>FaceNet</td>
<td>STI</td>
<td>80.30</td>
<td>31.50</td>
<td>24.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>86.74</td>
<td>31.51</td>
<td>24.32</td>
</tr>
</tbody>
</table>

Table 2. When the queries (from CALFW [72] and AgeDB [37]) 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 Tab. S2 and Tab. S3).
Figure 5. Figure in a similar format to that of Fig. 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. S5 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. 1).

Table 3. Our re-ranking (8×8 grid; APC) consistently improves the precision over Stage 1 alone (ST1) when identifying adversarial TALFW [73] images given an in-distribution LFW [65] gallery. The conclusions also carry over to other feature-weighting methods (more results in Tab. S5).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Method</th>
<th>P@1</th>
<th>RP</th>
<th>M@R</th>
</tr>
</thead>
<tbody>
<tr>
<td>TALFW [73] vs. LFW [65]</td>
<td>ArcFace</td>
<td>ST1</td>
<td>93.49</td>
<td>81.04</td>
<td>80.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>96.64</td>
<td>82.72</td>
<td>82.10</td>
</tr>
<tr>
<td>CosFace</td>
<td>ST1</td>
<td>96.49</td>
<td>83.57</td>
<td>82.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>99.07</td>
<td>85.48</td>
<td>85.03</td>
<td></td>
</tr>
<tr>
<td>FaceNet</td>
<td>ST1</td>
<td>95.33</td>
<td>79.24</td>
<td>78.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>97.26</td>
<td>80.33</td>
<td>79.39</td>
<td></td>
</tr>
</tbody>
</table>

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.

Experiment To generate augmented, masked images, we follow [10] 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 [6] (see Sec. S2). We train three models and report the mean and standard deviation (Tab. 4).

For a fair comparison, we evaluate the finetuned models and our no-training approach on the MLFW dataset [59], 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. 4; 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. 4; 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).

5. Related Work

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 [44]. Most prior methods propose to improve FI robustness by augmenting the training set of deep feature extractors with partially-occluded faces [23, 41, 57, 59, 63, 63]. Training on augmented, occluded data encourages models to rely more on local, dis-
Table 4. 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.

<table>
<thead>
<tr>
<th>ArcFace</th>
<th>Method</th>
<th>P@1</th>
<th>RP</th>
<th>M@R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-trained</td>
<td>(a) ST1</td>
<td>39.79</td>
<td>35.10</td>
<td>33.32</td>
</tr>
<tr>
<td></td>
<td>(b) Ours</td>
<td>48.23</td>
<td>41.43</td>
<td>39.71</td>
</tr>
<tr>
<td>Finetuned</td>
<td>(c) ST1</td>
<td>41.64± 0.16</td>
<td>34.67±0.24</td>
<td>32.66±0.25</td>
</tr>
<tr>
<td></td>
<td>(d) Ours</td>
<td>46.21± 0.27</td>
<td>38.65±0.26</td>
<td>36.73±0.26</td>
</tr>
</tbody>
</table>

Re-ranking in Face Identification Re-ranking is a popular 2-stage method for refining image retrieval results [69] in many domains, e.g., person re-identification [46], localization [51], or web image search [17]. In FI, Zhou et. al. [74] 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 [12]. Swearingen et. al. [55] 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 [74] or external knowledge [12]. Compared to face re-ranking [21, 42], 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 [45], its applications on deep convolutional features of images have been relatively under-explored. Zhang et al. [66, 67] 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 [24]). 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. [71] proposes DIML, which exhibits consistent improvement of ~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 [70]. 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 [60, 66, 67, 71], ours is the first to show the significant effectiveness of EMD on (1) occluded and adversarial OOD images; and (2) on face identification.

6. Discussion and Conclusion

Limitations 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. S1). Furthermore, here, we used EMD on two distributions of equal weights; however, the algorithm can be used for unequal-weight cases [16, 45], 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.

Acknowledgement We thank Qi Li, Peijie Chen, and Giang Nguyen for their feedback on manuscript. We also thank Chi Zhang, Wenliang Zhao, Chengrui Wang for releasing their DeepEMD, DIML, and MLFW code, respectively. AN was supported by the NSF Grant No. 1850117 and a donation from NaphCare Foundation.
References


Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Xiao-Xin Li, Dao-Qing Dai, Xiao-Fei Zhang, and Chuan-Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016. 6


Xiao-Xin Li, Dao-Qing Dai, Xiao-Fei Zhang, and Chuan-Xian Ren. Structured sparse error coding for face recognition with occlusion. IEEE transactions on image processing, 22(5):1889–1900, 2013. 8


Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 2, 12, 14


Chunlei Peng, Nannan Wang, Jie Li, and Xinbo Gao. Re-ranking high-dimensional deep local representation for nirvis face recognition. IEEE Transactions on Image Processing, 28(9):4553–4565, 2019. 8


Haibo Qiu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. End2end occluded face recognition by masking corrupted features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. 1, 2, 7, 8


20268


