Fast and Interpretable Face Identification for Out-Of-Distribution Data Using Vision Transformers

Hai Phan\textsuperscript{1} Cindy X. Le\textsuperscript{2} Vu Le\textsuperscript{3} Yihui He\textsuperscript{4} Anh “Totti” Nguyen\textsuperscript{1}
\{phai1204, anh.ng8, leducuvvietnam\}@gmail.com x12738@columbia.edu he2@alumni.cmu.edu
\textsuperscript{1}Auburn University \textsuperscript{2}Columbia University \textsuperscript{3}Phenikaa University \textsuperscript{4}Carnegie Mellon University

Abstract

Most face identification approaches employ a Siamese neural network to compare two images at the image embedding level. Yet, this technique can be subject to occlusion (e.g., faces with masks or sunglasses) and out-of-distribution data. DeepFace-EMD \cite{DeepFace-EMD} reaches state-of-the-art accuracy on out-of-distribution data by first comparing two images at the image level, and then at the patch level. Yet, its later patch-wise re-ranking stage admits a large $O(n^3 \log n)$ time complexity (for $n$ patches in an image) due to the optimal transport optimization.

In this paper, we propose a novel, 2-image Vision Transformers (ViTs) that compares two images at the patch level using cross attention. After training on 2M pairs of images on CASIA Webface \cite{CASIA Webface}, our model performs at a comparable accuracy as DeepFace-EMD on out-of-distribution data, yet at an inference speed more than twice as fast as DeepFace-EMD \cite{DeepFace-EMD}. In addition, via a human study, our model shows promising explainability through the visualization of cross-attention. We believe our work can inspire more explorations in using ViTs for face identification.

1. Introduction

Face identification (FI), the technology that enables automatic identification of individuals from photographs, is widely used in law enforcement \cite{Face ID in Law Enforcement, Face ID in Law Enforcement}, private businesses \cite{Face ID in Private Businesses}, smartphones \cite{Face ID in Smartphones}, and so on. 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, 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 \cite{DeepFace-EMD} or adversarially modified \cite{Adversarial Face Identification}.

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 \cite{Unlocking Devices}, and security checks \cite{Security Checks}, increasing people’s trust towards machine-generated results \cite{Machine-Generated Trust}, and lowering emergency response \cite{Emergency Response Time}. 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 \cite{End-User Operations} 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 \cite{Machine-User Interoperability, Machine-User Interoperability, Machine-User Interoperability}. Specifically, only a few studies have produced explanations for FI predictions \cite{DeepFace-EMD, Interpretable Face Identification} and none have evaluated the explanations from interpretable FI models on users.

In this work, we explore 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) ex-
plainability, we compare ViTs, CNNs, and EMD-based patch-wise re-ranking methods (Fig. 3) 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 [40]—a state-of-the-art approach to OOD face identification. In addition, our proposed model is more scalable as shown in Fig. 1, i.e., running over 2× faster in practice than DeepFace-EMD, which is slow due to the optimal transport optimization phase (Sec. 5.3).

- 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 (Sec. 5.4). We are the first to report that visual explanations improve end-user accuracy in face verification.

To our knowledge, our work is the first to (1) explore the design space of ViTs [16] for the FI problems on OOD data; (2) compare ViT-based and EMD-based image similarity approaches [40, 59, 60, 62]; and (3) study how visual explanations improve human accuracy in face verification.

2. Related Work

For face recognition, previous deep approaches typically adopt a CNN architecture (e.g. VGGNet [47], ResNet [23], etc.) as the backbone to extract deep face features and then use metric learning methods [12, 45] to classify identities. This design often achieves impressive results for in-distribution but fails on OOD data. A recent work, DeepFace-EMD [40], provided an Earth Mover’s Distance (EMD) distance to obtain cross-image information from CNN outputs, improving face OOD. Similar to DeepFace-EMD, we explore Transformers to exploit cross-attention information between inputs.

Out-of-distribution face identification. Identifying faces under occlusion [40, 42, 52] or adversarial changes [64] 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 [21, 39, 50, 52, 56, 56]. 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 [40]. An alternative technique for OOD face data is to reconstruct the missing pixels before performing FI [25, 32, 55, 57, 61, 66]. Yet, the de-occlusion process [8, 15, 20] 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 [12, 34, 45, 54]. Recent EMD-based image similarity work found that combining both image-level and patch-level similarity yields higher accuracy on in-distribution data [62] and OOD data [40, 60]. DeepFace-EMD [40] consistently outperforms traditional methods [12, 45, 54] 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 [16, 29, 49, 67], 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 [31] is leveraged to compare the similarity between intra-image patches [65] or between image-patches and text-tokens in image-text architectures [27, 31]. CrossViT [11] 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 [65], 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 [33, 53], 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. [6] 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. [27] use the similarity flow between text and image tokens as explanations for its similarity score. Chefer et al. [9, 10] 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 tech-
3. Method

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.

3.1. Problem Formulation

Similar to DeepFace-EMD [40], 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 sake of simplicity, we only consider data consisting of frontal faces without gestures.

3.2. Architecture: a two-Image Hybrid ViT

The overall architecture of the model is shown in Tab. 1, and Fig. 2. It takes in patch embeddings from a pre-trained CNN (ArcFace [12]). 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.

Given two input 2D face images $x_1, x_2$, we reshape them to have dimensions $\mathbb{R}^{H \times W \times C}$. The face embeddings $x_{p1}, x_{p2} \in \mathbb{R}^{P^2 \times D}$ 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 [12], we use $D = 512$ and $P = 8$.

We denote the learnable embeddings as $E$ and $E_{pos} \in \mathbb{R}^{(2 \times P^2 + 2) \times D}$, the two extra learnable embeddings as $X_{CLS}$ and $X_{SEP}$, and the intermediate layers of the Transformer encoder as $z$. $f_1$ and $f_2$ 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.

$$z_0 = [x_{CLS}E, x_{p1}E, x_{SEP}E, x_{p2}E] + E_{pos},$$

$$z'_l = \text{MSA}(\text{LayerNorm}(z_{l-1})), \quad l = 1 \ldots L$$

$$z_l = \text{MLP}(\text{LayerNorm}(z'_l)) + z'_l, \; l = 1 \ldots L$$

$$z_l \equiv [z_{CLS}, z^1_L, z_{SEP}, z^2_L], \; z^1_L, z^2_L \in \mathbb{R}^{P^2 \times D}$$

$$f_1 = \text{LayerNorm}(\text{Linear}_1(z^1_L))$$

$$f_2 = \text{LayerNorm}(\text{Linear}_2(z^2_L))$$

$$\text{loss} = \text{Arcface}_{\text{loss}}(f_1, f_2)$$

Position embeddings in vanilla Transformers [51] 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 $z'_l$ from a multihead-attention (MSA) layer are obtained through a combination of self and cross-attention processes. Previous ViT works [4, 11, 16, 28] 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 $z^1_L, z^2_L$ 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. 4. Similar to previous deep metric learning methods in face recognition [34, 45, 54], here we use the ArcFace as our loss function [12] to separate and learn cross-image margins to their corresponding labels.

3.3. Dataset

The model is trained on the CASIA Webface [58] dataset, containing 494,414 face images of 10,575 real-world identities, widely used for FI tasks such as [45]. We sample 2M pairs (1M positives and 1M negatives) consisting of all identities from the processed and clean CASIA Webface dataset.
### 3.4. Evaluation against various network structures

Here, we study six models with various architectures for face recognition, including a SNN with ArcFace [12], DeepFace-EMD [40], and Transformer ViTs, whose properties are summarized in Tab. 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 [16] (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 [40] (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. $8 \times 8$ in ResNet-18 [24]) are re-used to compute a feature vector through the linear layers which are deployed to ArcFace [12] 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. 1 and Sec. 3.4.

### 4. 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. 4.1), and (2) With cross-attention, 2-output linear vs. 1-output [CLS] (Sec. 4.2). In addition, we provide a study for how to select the depth and the head of Transformers...
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.

**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 [12] to classify identities.
improvement is even more significant when the accuracy increases by $+21.75$ points (Tab. 3; 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). 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. 5a, 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. Main Results

In Secs. 5.1 and 5.2, we experiment on different OOD query types including masks, sunglasses, and adversarial faces. Here, we select the best settings from ablation studies in Sec. 4 including depth of 1 and head of 1, 2, or 6. In Sec. 5.3, we show that our model has a faster time complexity compared with other layer types. Sec. 5.4 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 [40]. The models trained with settings mentioned in Sec. 4 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 [37,62]. For the details of these metrics, see Sec. S3.
5.1. Comparable accuracy

**Experiment.** We demonstrate our models for FI on two datasets: CALFW [63] and AgeDB [36]. 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 [40] consisting of masks and sunglasses.

**Results.** First, in ST1, 2-image (model H2L) achieves comparable accuracy with the original ArcFace [12]. In the AgeDB dataset, ST1's P@1 of model H2L improves around 42 points over model C on Mask (98.73% vs. 96.31%; Tab. 4) and Sunglasses (86.01% vs. 84.64%; Tab. 4), 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. 4) and achieves better results compared with DeepFaceEMD in sunglass images (ST2 on RP and M@R metrics in Tab. 4), verifying the boost performance in the 2-stage process.

5.2. Comparable robustness

**Experiment.** To illustrate the effectiveness of adversarial attacks, we run the experiment on the TALFW dataset [64]. 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. 4), 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. 4). These results show that these models (model H2L & D) are robust to adversarial images, which is a great challenge in computer vision [30, 38].

5.3. Faster inference time

<table>
<thead>
<tr>
<th>Layer type</th>
<th>Complexity per layer</th>
<th>Actual runtime (s)</th>
<th>Maximum path Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>C Convolutional</td>
<td>O(k \cdot n^2)</td>
<td>O(\log n)</td>
<td>(1)</td>
</tr>
<tr>
<td>V ViT, Self-Attention</td>
<td>O(n^2 \cdot d)</td>
<td>-</td>
<td>(1)</td>
</tr>
<tr>
<td>V Self-Attention (restricted)</td>
<td>O(1)</td>
<td>O(\log n)</td>
<td>(1)</td>
</tr>
<tr>
<td>H2L Hybrid-ViT</td>
<td>24.33</td>
<td>53.35</td>
<td>(1)</td>
</tr>
<tr>
<td>D DeepFace-EMD</td>
<td>43.70</td>
<td>55.57</td>
<td>(1)</td>
</tr>
</tbody>
</table>

Table 5. 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.

The run-time complexities of Model C, V, H2L, and D are shown in Tab. 5 and detailed in Sec. S4. Our Model H2L has a lower complexity, O(n^2), than that of DeepFaceEMD, 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. S1, Fig. 1). Moreover, in ST2, DeepFace-EMD is slow to solve EMD for higher dimension patch-wise similarity [40] 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 ~3× for 16,409 sunglass-query images in settings of 8 ×8 patches (see Tab. S1 for details). Therefore, model H2L is a good choice for more scalable architectures.

5.4. Better model explanation by human evaluation

As face identification systems in the real world are often customer-facing [7, 22, 26, 43, 44], 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. 6), 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 [40, 62], we use the cross-correlation method from [48] to generate similarity heatmaps for the CNNs and ViTs. This method produces a heatmap by taking 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 [40].

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
Figure 6. Comparison of face models’ explainability on LFW OOD domains. ViT-attn is visualized through the method of Chefer et al. [9]. 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.

Figure 7. 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 [48]; flow visualization in DeepFace-EMD [49]; 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%.

at the first layer of the ViT model uninformative to users (see Fig. 6; ViT-attn). Therefore, we use cross-correlation (CC) [48] to generate explanations for ViTs (Fig. 6; 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. S6 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. 7). In sum, consistent with the accuracy-based analysis in Sec. 5.1 & Sec. 5.2, our user study finds models with cross-image interaction (Model H2L and F) have higher explainability to users.

6. Discussion and Conclusion

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 remain the fastest due to efficient image embedding caching. Finally, visual explanations in cross-image interaction models greatly enhance lay-user face verification accuracy. We also conduct the inaugural study comparing state-of-the-art FI approaches based on 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 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 [5] for higher accuracy and incorporate sparsity into the attention mechanism of ViT for improved inference speed.

Acknowledgement. We thank Thang Pham, Giang Nguyen, Qi Li, Peijie Chen, Mohammad Reza Taesiri, and anonymous reviewers of WACV 2024 for valuable suggestions. Thanks to the 21 participants who dedicated their time to complete our user experiments. We are also grateful to AL EPSCoR GRSP 17 for funding HP. AN received support from NSF Grant No. 1850117 and a donation from Adobe Research and the NaphCare Foundation.
References


[26] Kashmir Hill. Flawed facial recognition leads to arrest and jail for new jersey man - the new york


