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

Hai Phan[1] Cindy X. Le[2] Vu Le[3] Yihui He[4] Anh "Totti" Nguyen[1]
{pthai1204, anh.ng8, leducvuvietnam}@gmail.com xl2738@columbia.edu he2@alumni.cmu.edu
[1]Auburn University [2]Columbia University [3]Phenikaa University [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 [40] 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 [58], 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 [40]. 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 [18, 43, 44], private businesses [2], smartphones [19], 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 [40] or adversarially modified [3, 64].

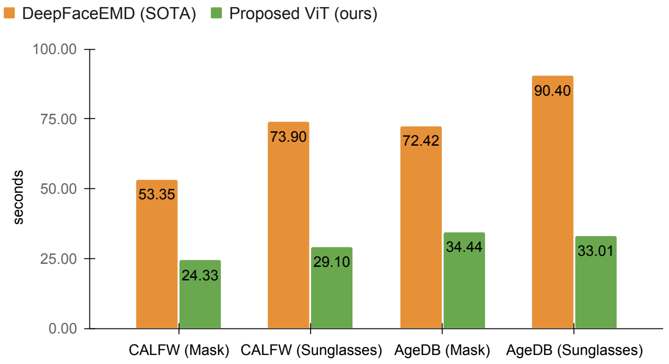


Figure 1. Actual running time in seconds (lower is better) for the re-ranking computation in face identification under occlusion. Our proposed model is at least two times faster than the state-of-the-art DeepFace-EMD [40] over all the datasets.

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 [19], and security checks [2], increasing people's trust towards machine-generated results [17], and lowering emergency response [35]. 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 [41] 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 [7, 22, 26, 43]. Specifically, only a few studies have produced explanations for FI predictions [40, 48] 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) explainability, we compare ViTs, CNNs, and EMD-based patch-wise re-ranking methods (Fig. 3) and find that: [1]

- 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 [51] 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-

[1] Code, demo and data are available at https://github.com/anguyen8/face-vit

nique proposed by [48].

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 scope of this paper, 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.

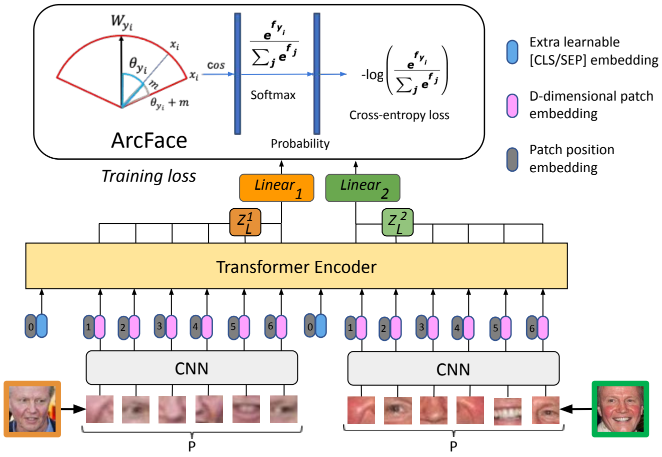


Figure 2. The architecture of the proposed ViT-based Model H2L.

Face embeddings. For a zero-shot face problem, deep metric learning works efficiently [34,45,54]. Besides *[CLS]* (classification tokens) [13, 16] 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. 2 for details).

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 [12], 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.

$$\mathbf{z}_0 = [\mathbf{x}_{CLS}\mathbf{E}, \mathbf{x}_{p1}\mathbf{E}, \mathbf{x}_{SEP}\mathbf{E}, \mathbf{x}_{p2}\mathbf{E}] + \mathbf{E}_{pos}, \quad (1)$$

$$\mathbf{z}_l^{'} = \text{MSA}(\text{LayerNorm}(\mathbf{z}_{l-1})), \quad l = 1 \ldots L \quad (2)$$

$$\mathbf{z}_l = \text{MLP}(\text{LayerNorm}(\mathbf{z}_l^{'})) + \mathbf{z}_l^{'}, \; l = 1 \ldots L \quad (3)$$

$$\mathbf{z}_l \equiv [\mathbf{z}_{CLS}, \mathbf{z}_L^1, \mathbf{z}_{SEP}, \mathbf{z}_L^2], \;\; \mathbf{z}_L^1, \mathbf{z}_L^2 \in \mathbb{R}^{P^2 \times D} \quad (4)$$

$$\mathbf{f}_1 = \text{LayerNorm}(\text{Linear}_1(\mathbf{z}_L^1)) \quad (5)$$

$$\mathbf{f}_2 = \text{LayerNorm}(\text{Linear}_2(\mathbf{z}_L^2)) \quad (6)$$

$$\text{loss} = \text{Arcface\_loss}(\mathbf{f}_1, \mathbf{f}_2) \quad (7)$$

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 $\mathbf{z}_l^{'}$ from a multi-head-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 $\mathbf{z}_L^1, \mathbf{z}_L^2$ 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. 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.

Name	Architecture	Patch Embedding	Input	Transformer output	**Inter**-image, Image-wise comparison	**Intra**-image, patch-wise comparison	**Inter**-image, patch-wise comparison
C	CNN [12]	CNN [1]	1-image	1 feature	✓	Local (CNN-based)	✗
V	ViT [16]	*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 [40]	CNN	2-image	2 features	✓($\alpha = 0.3$)	Local (CNN-based)	✓($\alpha = 0.7$)

Table 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.

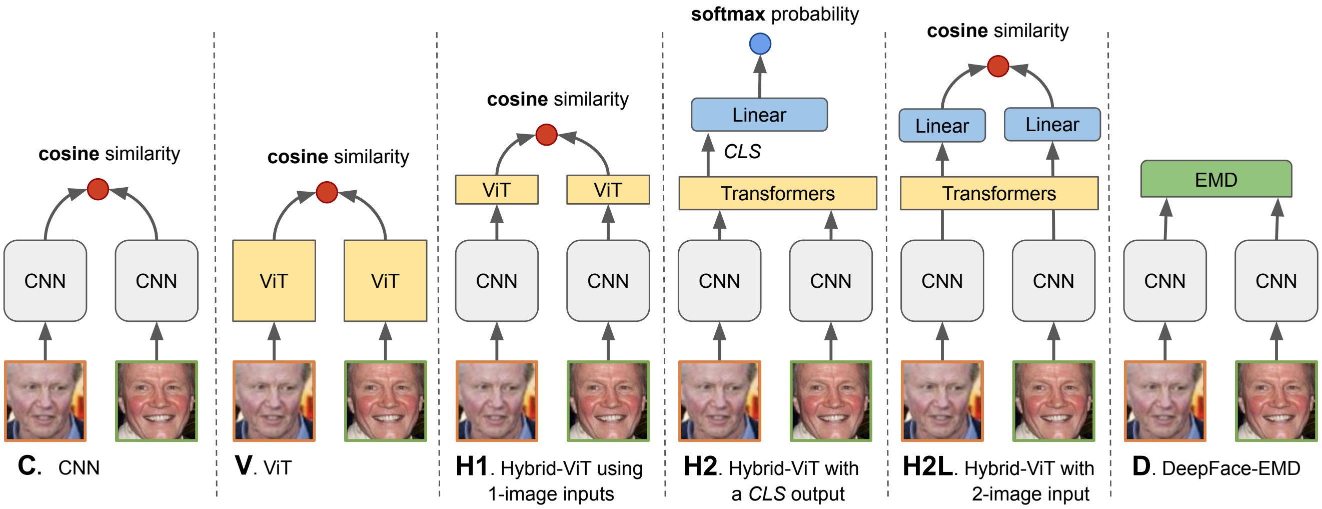


Figure 3. The architecture of the six networks evaluated in this work including our proposed H2L.

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.* 8×8 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

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 2. 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.

(Sec. S5).

Datasets. We run face verification experiments on two datasets: the in-distribution LFW [58] and the masked-face-occlusion MLFW [52]. 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 [12]. Images are aligned and cropped to 128 × 128 by the MTCNN algorithm [45]. 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 [58].

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 [12], hyper-parameters are mentioned in Sec. S2. All experiments are run on eight 40GB A100 SXM GPUs.

4.1. 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.

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.

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

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. 2). 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 similarity information than the 1-image model for OOD distribution on MLFW (Tab. 2; Model H2L - 73% vs. C-70.75%, H1-57.92%, and V-57.62%).

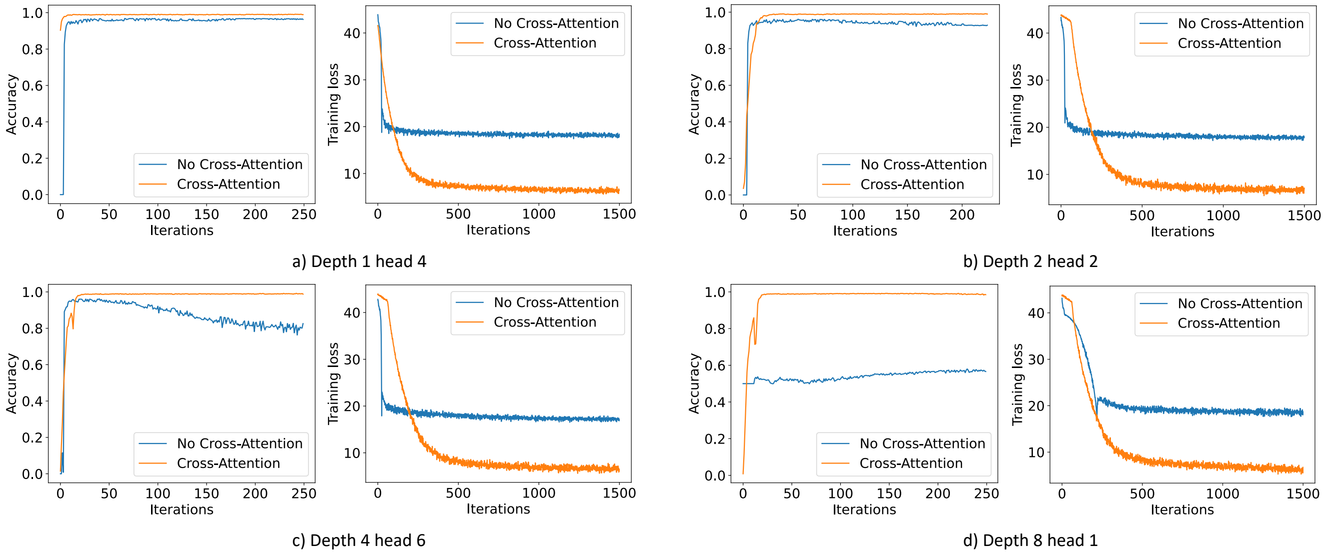
Second, interestingly, we find that the hybrid models (H1 & H2L) can achieve higher precision with a depth of only 1, *i.e.* adding an efficient shallow layer to Transformers can improve performance (*e.g.* on LFW, 99.28% H2L vs. 98.02 % of H1). We deduce the same statement when comparing it with the ViT model (V). In contrast, the 1-image no-cross-attention model has worse performance with the in-distribution LFW (see Fig. 4) and the OOD MLFW (Tab. 2). With a higher depth of 8, model H1 becomes worse in LFW (Tab. 2 H2L-99.22% vs. H1-58.00%)

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

The previous Transformer-based FI works [14, 16] 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. 3), 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. 3; from 90.45% of H2 to 99.22% of H2L). In the out-of-distribution masked-face image (MLFW) datasets, the



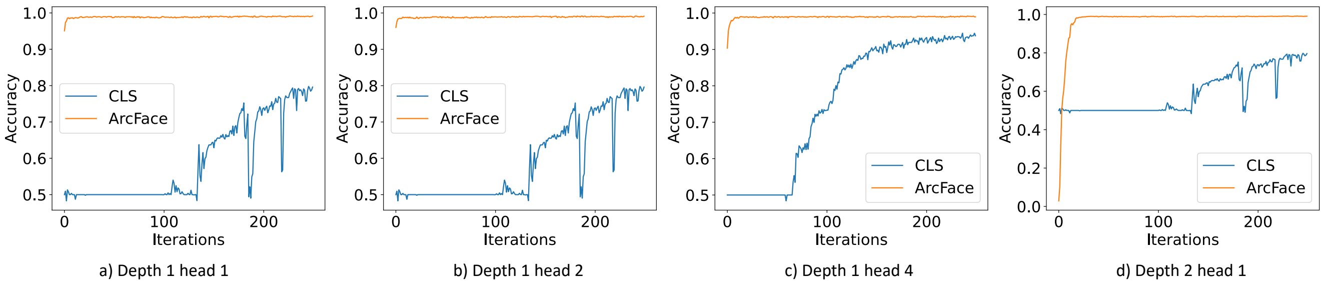
a) Depth 1 head 4

b) Depth 2 head 2

c) Depth 4 head 6

d) Depth 8 head 1

Figure 4. Comparison in accuracy and convergence between training **H1** (No-cross-attention) vs. **H2L** (Cross-attention) architectures on LFW [58]. For different network settings, 2-input-image achieves better accuracy and more stable training when leveraging patch-wise cross-image attention.



a) Depth 1 head 1

b) Depth 1 head 2

c) Depth 1 head 4

d) Depth 2 head 1

Figure 5. 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.

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.

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 4. Face occlusions and adversarial images. Model H2L achieves comparable accuracy on the OOD of CALFW and AgeDB compared to CNN and DeepFace-EMD [40].

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 +2 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 (models H2L & D) are robust to adversarial images, which is a grand challenge in computer vision [30, 38].

5.3. Faster inference time

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 [40]	$O(k \cdot n \cdot d^2 + n^3 \cdot \log n)$ [46]	53.35	$O(1)$

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 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. 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 $\sim$ **3**$\times$ 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

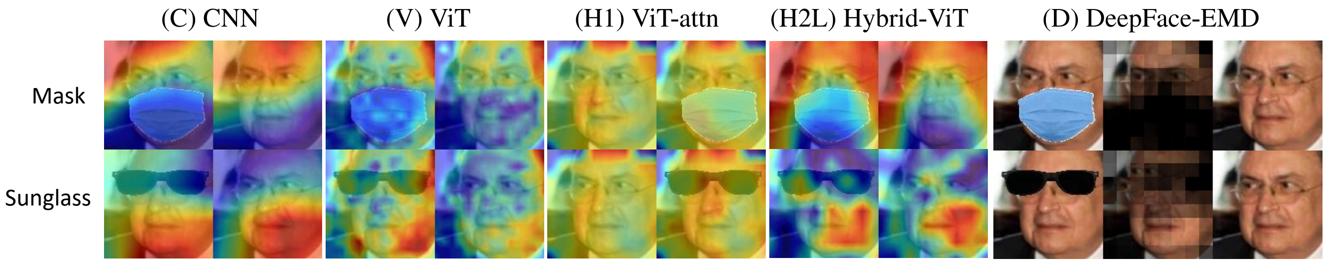


Figure 6. Comparison of face models' explainability on **LFW** OOD domains. ViT-attn is visualized through the method of Chefer et al. [9]. Our proposed H2L can highlight the important area in images (*e.g*. eyes, mouth, etc.) and remove occluded areas (*e.g*. mask and sunglasses). In contrast, Model V contains noisy heatmaps and H1 does not provide any interpretable clues of how two faces match.

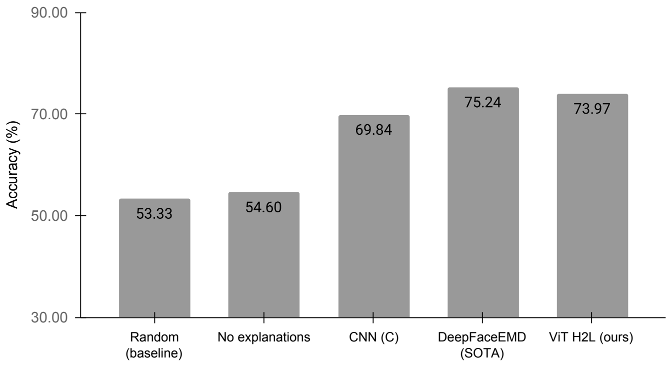


Figure 7. Human explainability across various networks. The mean and standard deviation of the accuracy of 21 users when presented with 4 explanations: Cross-correlation (CC) method on CNNs [48]; flow visualization in DeepFace-EMD [40]; CC on 2-image Hybrid-ViT; and a baseline of no explanations. The 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

[1] ronghuaiyang/arcface-pytorch. https://github.com/ronghuaiyang/arcface-pytorch. (Accessed on 11/16/2021). 4

[2] Satariano Adam and Hill 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). 1

[3] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016. 1

[4] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. *ICLR 2022*, 2022. 3

[5] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022. 8

[6] Samuel Black, Abby Stylianou, Robert Pless, and Richard Souvenir. Visualizing paired image similarity in transformer networks. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1534–1543, 2022. 2

[7] Nann Melissa 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). 1, 7

[8] Jiancheng Cai, Hu Han, Jiyun Cui, Jie Chen, Li Liu, and S Kevin Zhou. Semi-supervised natural face de-occlusion. *IEEE Transactions on Information Forensics and Security*, 16:1044–1057, 2020. 2

[9] Hila Chefer, Shir Gur, and Lior 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. 2, 8

[10] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 2

[11] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar 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. 2, 3

[12] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2, 3, 4, 5, 7, 12

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 5

[15] Jiayuan Dong, Liyan Zhang, Hanwang Zhang, and Weichen 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. 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 3, 4, 5

[17] Emir Efendić, Philippe PFM Van de Calseyde, and Anthony M Evans. Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, 157:103–114, 2020. 1

[18] Alex 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). 1

[19] Chaim 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). 1

[20] Shiming Ge, Chenyu Li, Shengwei Zhao, and Dan 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. 2

[21] Jianzhu Guo, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Face synthesis for eyeglass-robust face recognition. In *Chinese Conference on biometric recognition*, pages 275–284. Springer, 2018. 2

[22] Drew 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). 1, 7

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 4, 12

[25] Ran He, Wei-Shi Zheng, Bao-Gang Hu, and Xiang-Wei Kong. A regularized correntropy framework for robust pattern recognition. *Neural computation*, 23(8):2074–2100, 2011. 2

[26] Kashmir 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). 1, 7

[27] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong 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. 3

[29] Rajat Koner, Poulami Sinhamahapatra, Karsten Roscher, Stephan Günnemann, and Volker Tresp. Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976*, 2021. 2

[30] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016. 7

[31] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[32] 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. 2

[33] Shengcai Liao and Ling Shao. Transmatcher: Deep image matching through transformers for generalizable person reidentification. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[34] 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, 3

[35] Sato 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). 1

[36] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos 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. 7

[37] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *ECCV*, pages 681–699. Springer, 2020. 6, 13

[38] Anh Nguyen, Jason Yosinski, and Jeff 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. 7

[39] Elad Osherov and Michael Lindenbaum. Increasing cnn robustness to occlusions by reducing filter support. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 550–561, 2017. 2

[40] Hai Phan and Anh 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. 1, 2, 3, 4, 6, 7, 8, 13, 16

[41] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami 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. 1

[42] 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. 2

[43] Tate 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). 1, 7

[44] Mia 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). 1, 7

[45] Florian Schroff, Dmitry Kalenichenko, and James 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. 2, 3, 5

[46] Sameer Shirdhonkar and David W. Jacobs. Approximate earth mover's distance in linear time. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 7

[47] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 2

[48] Abby Stylianou, Richard Souvenir, and Robert Pless. Visualizing deep similarity networks. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 2029–2037. IEEE, 2019. 1, 3, 7, 8, 13

[49] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In Marina Meila and Tong 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. 2

[50] Daniel Sáez Trigueros, Li Meng, and Margaret Hartnett. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. *Image and Vision Computing*, 79:99–108, 2018. 2

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3, 13

[52] Chengrui Wang, Han Fang, Yaoyao Zhong, and Weihong Deng. Mlfw: A database for face recognition on masked faces. *arXiv preprint arXiv:2109.05804*, 2021. 2, 5

[53] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. Nformer: Robust person re-identification with neighbor transformer. *arXiv preprint arXiv:2204.09331*, 2022. 2

[54] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018. 2, 3

[55] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2008. 2

[56] Xiang Xu, Nikolaos Sarafianos, and Ioannis 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. 2

[57] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. Robust sparse coding for face recognition. In *CVPR 2011*, pages 625–632. IEEE, 2011. 2

[58] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 1, 3, 5, 6, 12

[59] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua 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. 2

[60] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua 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. 2

[61] F. Zhao, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan. Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE Transactions on Image Processing*, 27:778–790, 2018. 2

[62] Wenliang Zhao, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Towards interpretable deep metric learning with structural matching. In *ICCV*, 2021. 2, 6, 7

[63] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017. 7

[64] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2020. 1, 2, 7

[65] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021. 2

[66] Zihan Zhou, Andrew Wagner, Hossein Mobahi, John Wright, and Yi Ma. Face recognition with contiguous occlusion using markov random fields. In *2009 IEEE 12th international conference on computer vision*, pages 1050–1057. IEEE, 2009. 2

[67] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2