JAN 4-8 WACV 2024 WAIKOLOA HAWAII

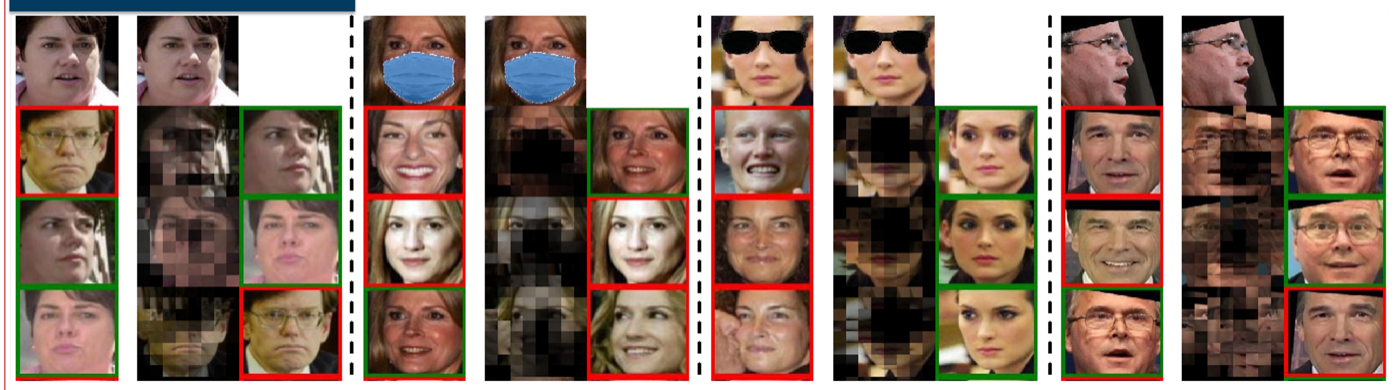
Face-ViT's code and demo

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

Hai Phan[1], Cindy Le[2], Vu Le[3], Yihui He[4], Anh Nguyen[1]

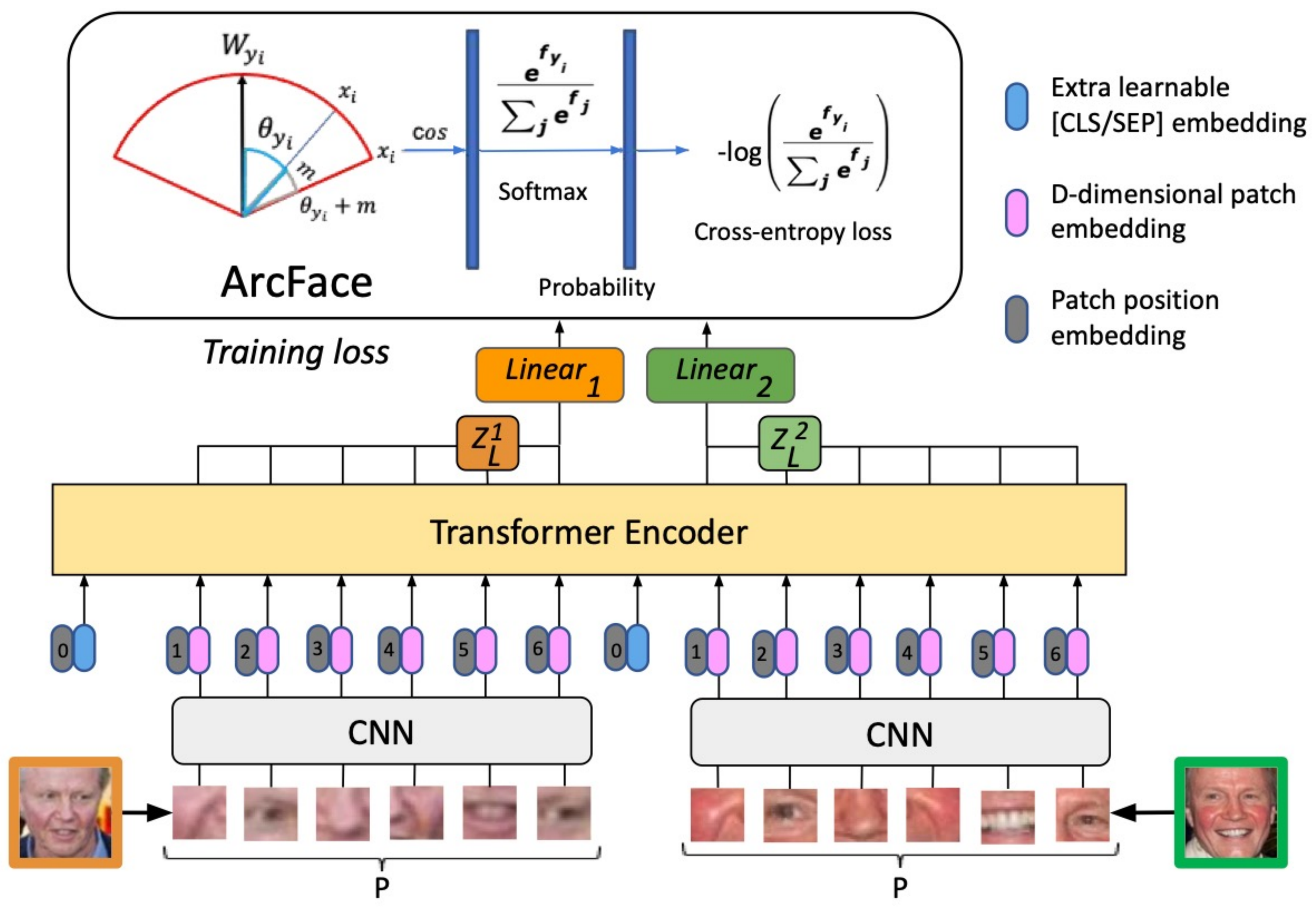
[1]Auburn University. [2]Columbia University. [3]Phenikaa University. [4]Carnegie Mellon University

Summary



- **Face Identification** (FI) is today is behind the answers to many life-critical questions (e.g. who are you to receive unemployment benefits or boarding to planes?, etc.)
- Current face verification accuracy **may notoriously drop significantly** (from **99.38%** to **81.12%** on LFW) given an occluded queries or adversarial queries.
- We propose to evaluate **the performance of SOTA facial feature extractors** (e.g. ArcFace, CosFace, etc.) on OOD FI test. The main task is to recognize the person in a query image given a gallery of know faces. The evaluation is on 3 metrics: **P@1**, **RP**, and **M@R**.

Methods



- **Stage 1:** Ranking gallery images based on their pair-wise cosine sim.
- **Stage 2 (Re-ranking)**: re-rank top-k (e.g. **100**) candidates from Stage 1 by computing patch-wise similarity for an image pair using EMD.
- **Goal**: Find **the optimal flows** between query and gallery images to select important features networks used for matching.

Formulation

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

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

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

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

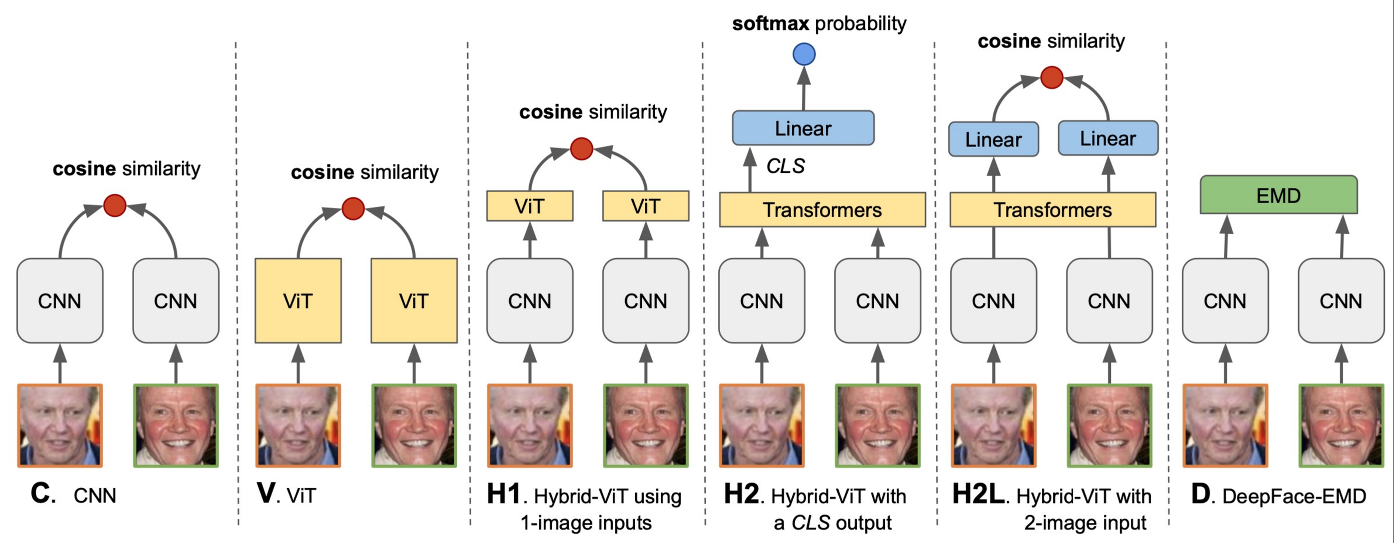
$$\mathbf{f}_1 = \mathrm{LayerNorm}(\mathrm{Linear}_1(\mathbf{z}_L^1)) \quad (5)$$

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

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

Ablation Study

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)	✓($\alpha = 0.7$)

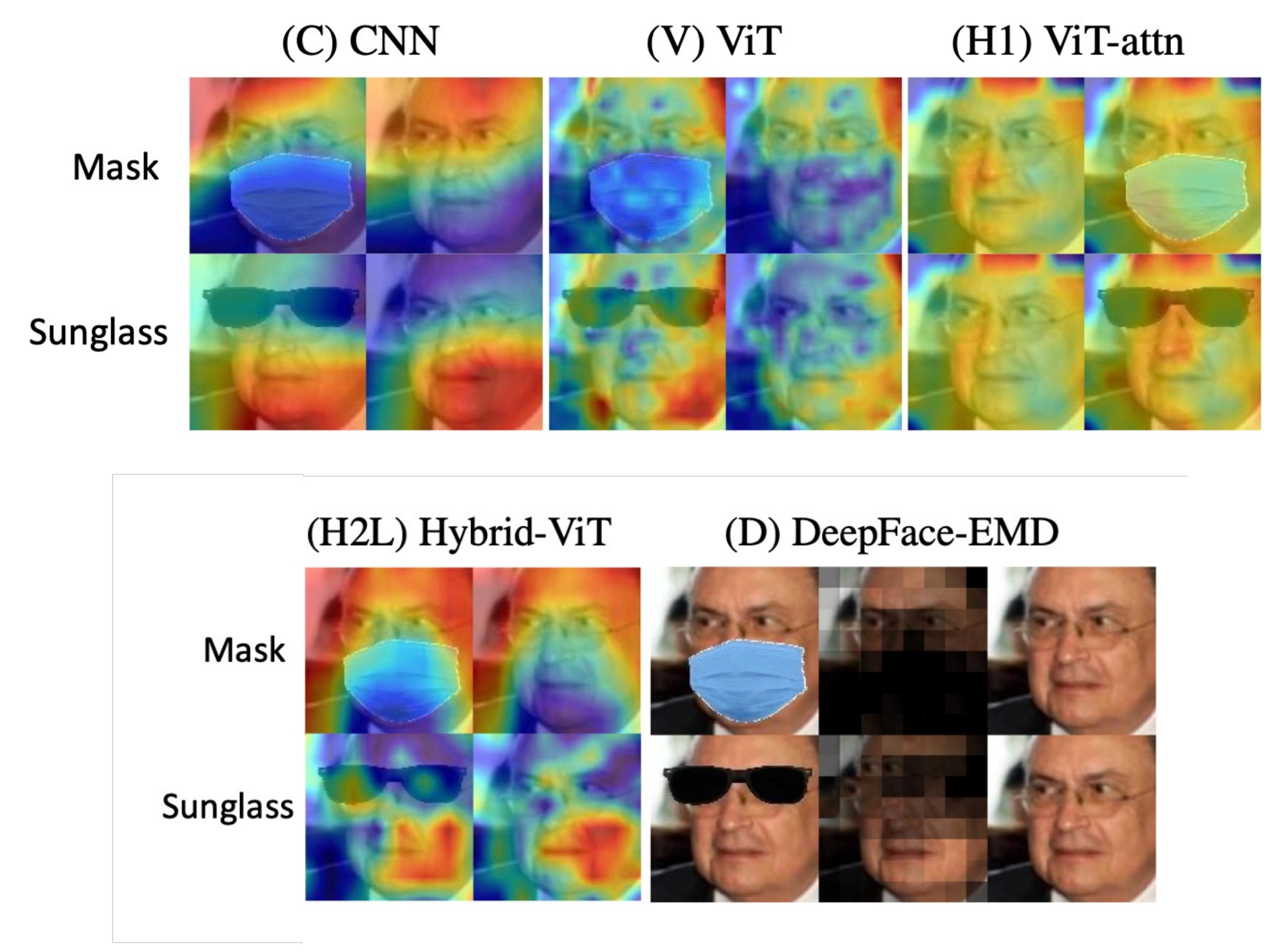


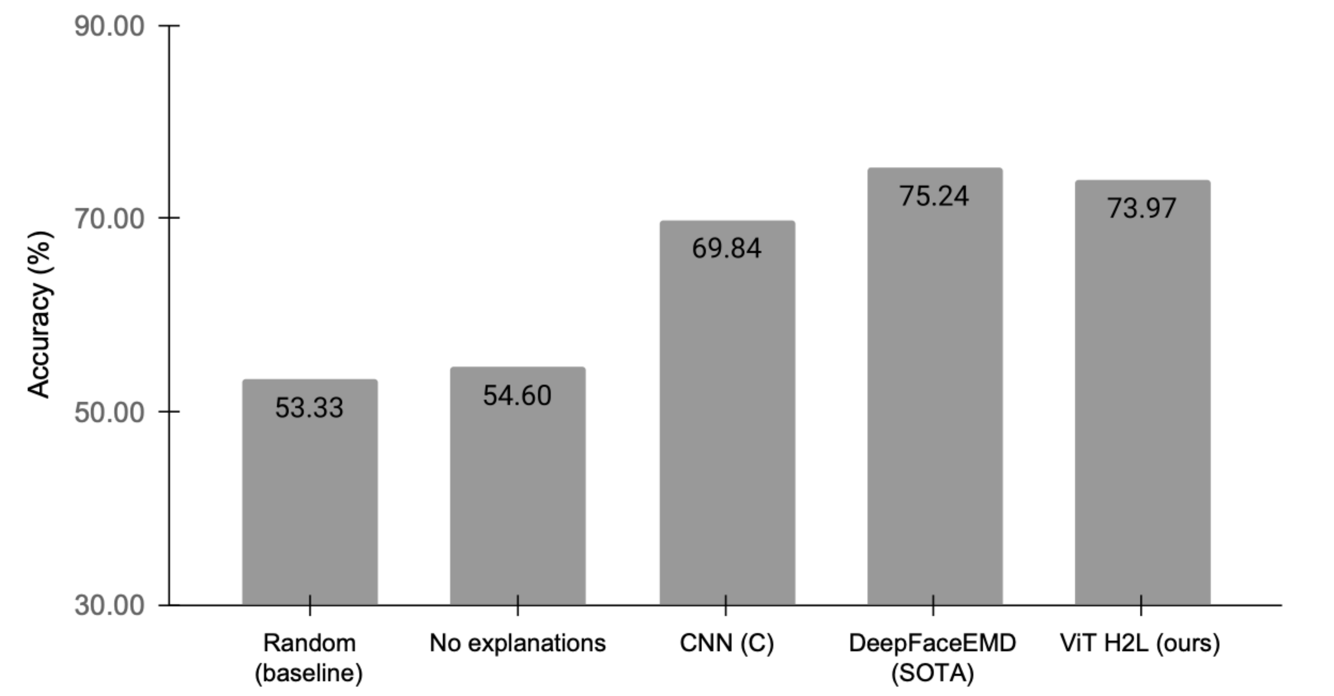
Accuracy

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

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

Explainability





Time Complexity

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	$O(1)$

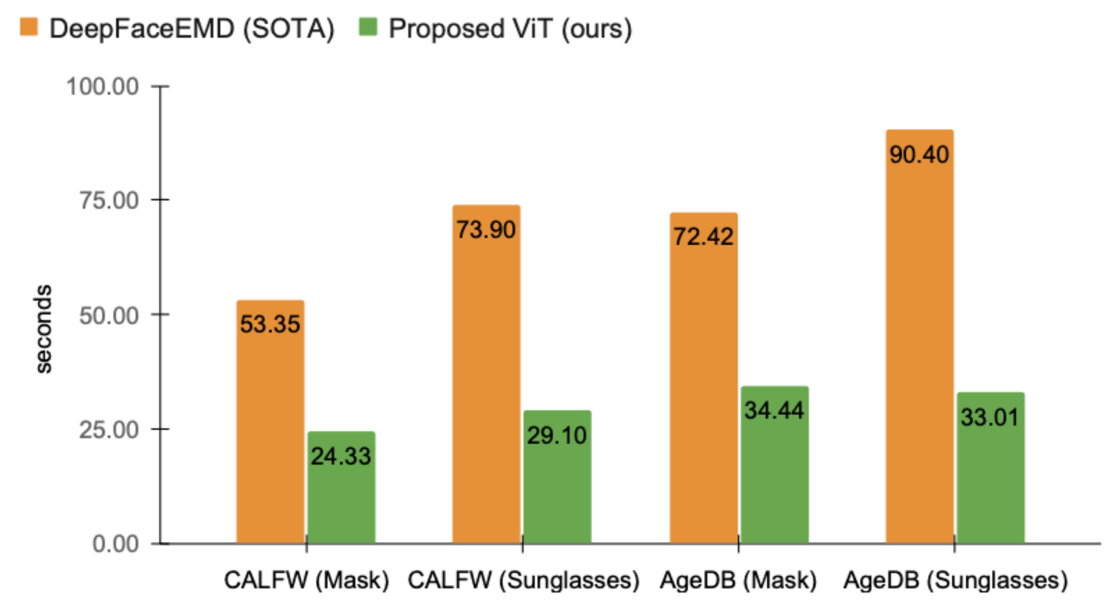


Figure 1. Actual running time in seconds (lower is better) for the re-ranking computation in face identification under occlusion. Our proposed model is at least two times faster than the state-of-the-art DeepFace-EMD [40] over all the datasets.

User Study