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

S1. Pre-trained models

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

- ArcFace [19]: https://github.com/ronghuaiyang/arcface-pytorch
- FaceNet [47]: https://github.com/timesler/facenet-pytorch
- CosFace [61]: https://github.com/MuggleWang/CosFace_pytorch

These ArcFace, FaceNet, and CosFace models were trained on dataset CASIA Webface [65], VGGFace2 [15], and CASIA Webface [65], respectively.

Architectures The network architectures are provided here:

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

Image-level embeddings for Ranking We use these layers to extract the image embeddings for stage 1, *i.e.*, ranking images based on the cosine similarity between each pair of (query image, gallery image).

- Arcface: layer bn5 (see code), which is the 512-output, last BatchNorm linear layer of ArcFace (a modified ResNet-18 [24]).
- FaceNet: layer last_bn (see code), which is the 512-output, last BatchNorm linear layer of FaceNet (an Inception-ResNet-v1 [56]).
- CosFace: layer fc (see code), which is the 512-output, last linear layer of the 20-layer SphereFace architecture [33].

Patch-level embeddings for Re-ranking We use the following layers to extract the spatial feature maps (*i.e*. embeddings $\{q_i\}$) for the patches:

- ArcFace: layer dropout (see code). Spatial dimension: 8×8.
- FaceNet: layer block8 (see code) Spatial dimension: 3×3.
- CosFace: layer layer4 (see code). Spatial dimension: 6×7.

S2. Finetuning hyperparameters

We describe here the hyperparameters used for finetuning ArcFace on our CASIA dataset augmented with masked images (see Fig. S6 for some samples).

- Training on $907,459$ facial images (masks and non-masks).
- Number of epochs is 12.
- Optimizer: SGD.
- Weight decay: $5e^{-4}$
- Learning rate: 0.001
- Margin: $m = 0.5$
- Feature scale: $s = 30.0$

See details in the published code base: code

S3. Flow visualization

We use the same visualization technique as in DeepEMD to generate the flow visualization showing the correspondence between two images (see the flow visualization in Fig. 1 or Fig. S2). Given a pair of embeddings from query and gallery images, EMD computes the optimal flows (see Eq. (1) for details). That is, given a 8×8 grid, a given patch embedding q_i in the query has 64 flow values $\{f_{ij}\}$ where $j \in \{1, 2, ..., 64\}$. In the location of patch q_i in the query image, we show the corresponding highest-flow patch g_k, *i.e.* k is the index of the gallery patch of highest flow $f_{i,k} = \max(f_{i,1}, f_{i,2}, ..., f_{i,64})$. For displaying, we normalize a flow value $f_{i,k}$ over all 64 flow values (each for a patch $i \in \{1, 2, ..., 64\}$) via:

$$f = \frac{f - \min(f)}{\max(f) - \min(f)} \tag{10}$$

See Fig. S4, Fig. S5, and Fig. 5 for example flow visualizations.

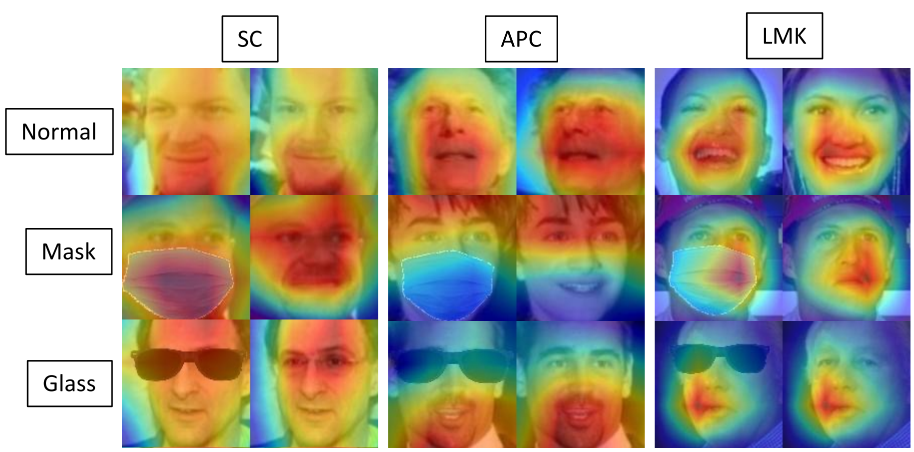


Figure S1. The feature-weighting heatmaps using SC, APC, and LMK for random pairs of faces across three input types (normal faces, and faces with masks and sunglasses). Here, we use ArcFace [19] and an 4×4 grid (average pooling result from 8×8). SC heatmaps often cover the entire face including the occluded region. APC tend to assign low importance to occlusion and the corresponding region in the unoccluded image (see blue areas in APC). LMK results in a heatmap that covers the middle area of a face. Best view in color.

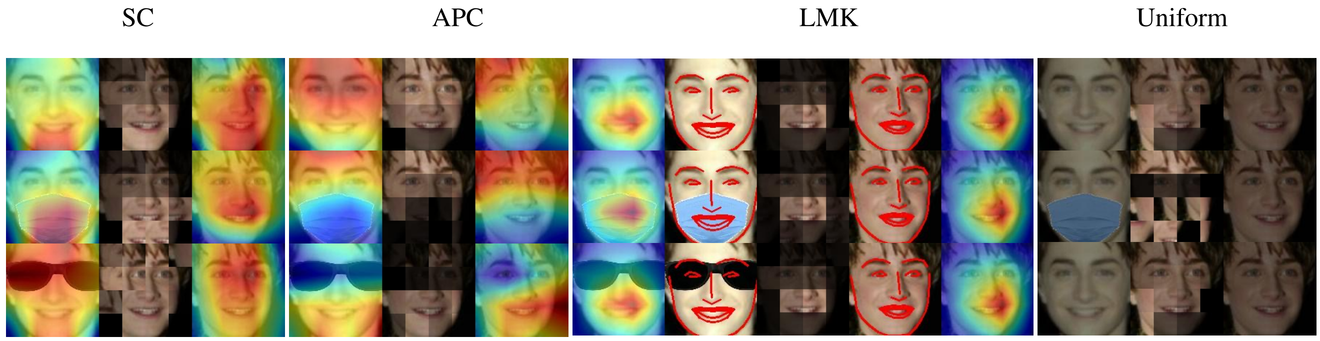


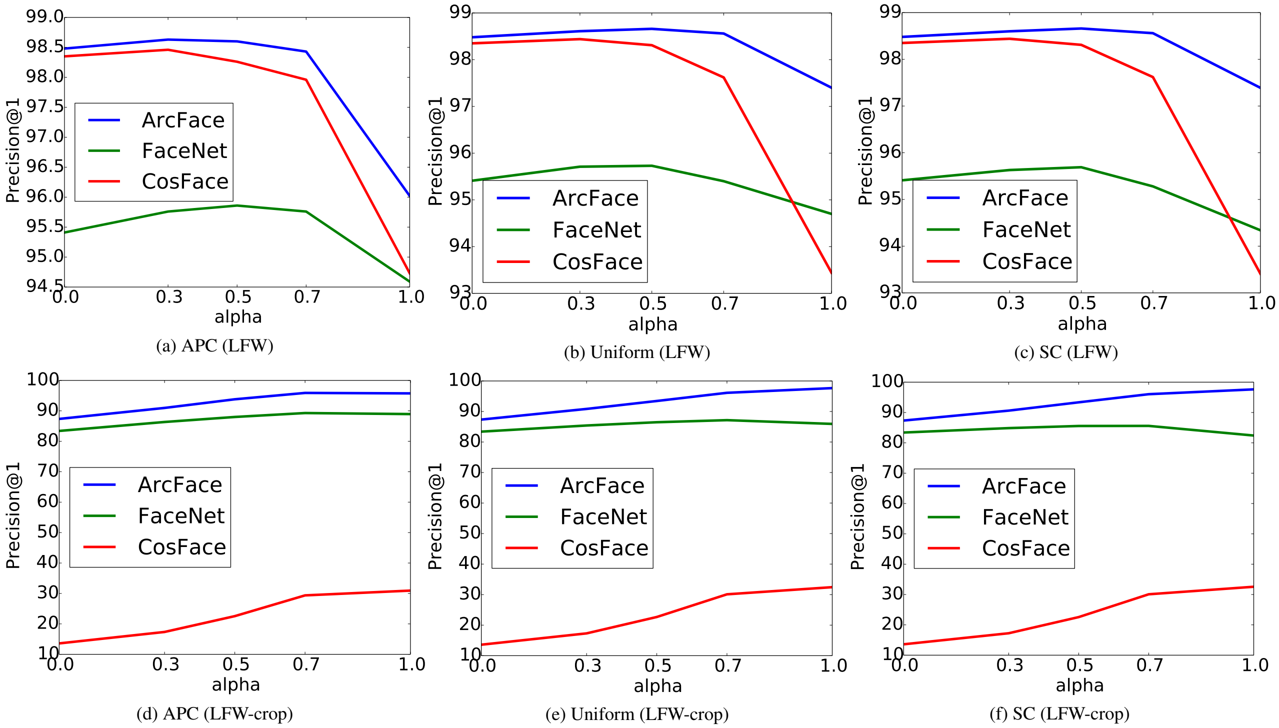
Figure S2. Given a pair of images, after the features are weighted (heatmaps; red corresponds to 1 and blue corresponds to 0 importance weight), EMD computes an optimal matching or "transport" plan. The middle flow image shows the one-to-one correspondence following the format in [66] (see also description in Sec. S3). That is, intuitively, the flow visualization shows the reconstruction of the left image, using the nearest patches (i.e. highest flow) from the right image. Here, we use ArcFace and a 4$\times$ patch size (*i.e*. computing the EMD between two sets of 16 patch-embeddings). Darker patches correspond to smaller flow values. How EMD computes facial patch-wise similarity differs across different feature weighting techniques (SC, APC, LMK, and Uniform).

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

Table S1. Comparison of performing patch-wise EMD ranking at Stage 1 vs. our proposed 2-stage FI approach (*i.e*. cosine similarity ranking in Stage 1 and patch-wise EMD re-ranking in Stage 2). In both cases, EMD uses 8$\times$8 patches. EMD at Stage 1 is the method of using EMD to rank images directly (instead of the regular cosine similarity) and there is no Stage 2 (re-ranking). For our method, we choose the same setup of $\alpha = 0.7$. Our 2-stage approach does not only outperform using EMD at Stage 1 but is also $\sim$2-4 $\times$ faster. The run time is the total for all **13,214 queries** for both (a) and (b). The result supports our choice of performing EMD in Stage 2 instead of Stage 1.

S4. Additional Results: Face Verification on MLFW

In the main text, we find that DeepFace-EMD is effective in face *identification* given many types of OOD images. Here, we also evaluate DeepFace-EMD for face *verification* of MLFW [59], a recent benchmark that consists of masked LFW faces. As in common verification setups of LFW [33, 47, 59], given pairs of face images and their similarity scores predicted by a verification system, we find the optimal threshold that yields the best accuracy. Here, we follow the setup in [59] to enable a fair comparison. First of all, we reproduce Table 3 in [59], which evaluate face verification accuracy on 6,000 pair of MLFW images. Then, we run our DeepFace-EMD distance function (Eq. 9). We found that using our proposed distance consistently improves on face *verification* for all three PyTorch models in [59]. Interestingly, with DeepFace-EMD, **we obtained a state-of-the-art result** (91.17%) on MLFW (see Tab. S6).



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

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

Figure S3. The P@1 of our 2-stage FI when sweeping across $\alpha \in \{0, 0.3, 0.5, 0.7, 1.0\}$ for linearly combining EMD and cosine distance on LFW (top row; a–c) and LFW-crop images (bottom row; d–f) of all feature weighting (APC, Uniform, and SC).

S5. Additional ablation studies: 3D Facial Alignment vs. MTCNN

The reason we used the 3D alignment pre-processing instead of the default MTCNN pre-processing [68] of the three models was because for ArcFace, the 3D alignment actually resulted in better P@1, RP, and M@R for both our baselines and DeepFace-EMD (*e.g.* +3.35% on MLFW). For FaceNet, the 3D alignment did yield worse performance compared to MTCNN. However, we confirm that our conclusions that **DeepFace-EMD improves FI on the reported datasets regardless of the pre-processing choice**. See Tab. S7 for details.

Dataset	Model	Method	P@1	RP	M@R
CALFW (Mask)	ArcFace	Stage 1	96.81	53.13	51.70
		APC	**99.92**	**57.27**	**56.33**
		Uniform	**99.92**	**57.28**	**56.24**
		SC	**99.92**	**57.13**	**56.06**
	CosFace	Stage 1	98.54	43.46	41.20
		SC	**99.96**	**59.87**	**58.93**
		Uniform	**99.96**	**59.86**	**58.91**
		APC	**99.96**	**59.85**	**58.87**
	FaceNet	Stage 1	77.63	39.74	36.93
		APC	**96.67**	**45.87**	**44.53**
		Uniform	**94.23**	**43.90**	**42.33**
		SC	**90.80**	**42.85**	**40.95**
CALFW (Sunglass)	ArcFace	Stage 1	51.11	29.38	26.73
		Uniform	**55.80**	**31.50**	**28.60**
		APC	**54.95**	**30.66**	**27.74**
		SC	**55.45**	**31.42**	**28.49**
	CosFace	Stage 1	45.20	25.93	22.78
		Uniform	**50.28**	**27.23**	**24.40**
		APC	**49.67**	**26.98**	**24.12**
		SC	**50.24**	**27.22**	**24.38**
	FaceNet	Stage 1	21.68	13.70	10.89
		APC	**25.07**	**15.04**	**12.16**
		Uniform	**25.08**	**14.97**	**12.21**
		SC	**24.38**	**14.58**	**11.88**
CALFW (Crop)	ArcFace	Stage 1	79.13	43.46	41.20
		Uniform	**94.04**	**49.57**	**48.15**
		APC	**92.57**	**47.17**	**45.68**
		SC	**93.76**	**49.51**	**48.05**
	CosFace	Stage 1	10.99	6.45	5.43
		SC	**27.42**	**12.68**	**11.59**
		Uniform	**27.43**	**12.66**	**11.58**
		APC	**25.99**	**12.35**	**11.13**
	FaceNet	Stage 1	79.47	44.40	41.99
		APC	**85.71**	**45.91**	**43.83**
		Uniform	**83.92**	**45.22**	**43.04**
		SC	**82.33**	**44.54**	**42.26**

Table S2. Our 2-stage method for all feature weighting methods (APC, SC, and Uniform) for face occlusions (*e.g*. mask, sunglass, and crop) is substantially more robust to the Stage 1 alone baseline (ST1) on CALFW [72].

Dataset	Model	Method	P@1	RP	M@R
AgeDB (Mask)	ArcFace	Stage 1	96.15	39.22	30.41
		APC	**99.84**	**39.22**	**33.18**
		Uniform	**99.82**	**39.23**	**32.94**
		SC	**99.82**	**39.12**	**32.77**
	CosFace	Stage 1	98.31	38.17	31.57
		APC	**99.95**	**39.70**	**33.68**
		Uniform	**99.95**	**39.61**	**33.60**
		SC	**99.95**	**39.63**	**33.62**
	FaceNet	Stage 1	75.99	22.28	14.95
		APC	**96.53**	**24.25**	**17.49**
		Uniform	**93.99**	**22.55**	**15.68**
		SC	**90.60**	**22.14**	**15.13**
AgeDB (Sunglass)	ArcFace	Stage 1	84.64	51.16	44.99
		Uniform	**88.06**	**51.17**	**45.24**
		APC	**87.06**	**50.40**	**44.27**
		SC	**87.96**	**51.16**	**45.22**
	CosFace	Stage 1	68.93	34.90	27.30
		APC	**75.97**	**35.54**	**28.12**
		Uniform	**74.85**	**35.33**	**27.79**
		SC	**74.82**	**35.33**	**27.79**
	FaceNet	Stage 1	56.77	27.92	20.00
		APC	**61.21**	**28.98**	**21.11**
		Uniform	**61.64**	**28.62**	**20.94**
		SC	**61.27**	**28.44**	**20.76**
AgeDB (Crop)	ArcFace	Stage 1	79.92	32.66	26.19
		Uniform	**94.18**	**34.81**	**28.80**
		APC	**92.92**	**32.93**	**26.60**
		SC	**94.03**	**34.83**	**28.80**
	CosFace	Stage 1	10.11	4.23	2.18
		SC	**21.00**	**5.02**	**2.89**
		Uniform	**20.96**	**5.02**	**2.88**
		APC	**19.58**	**4.95**	**2.76**
	FaceNet	Stage 1	80.80	31.50	24.27
		APC	**86.74**	**31.51**	**24.32**
		Uniform	**84.93**	**30.87**	**23.68**
		SC	**83.29**	**30.51**	**23.24**

Table S3. Our 2-stage method for all feature weighting methods (APC, SC, and Uniform) for face occlusions (*e.g*. mask, sunglass, and crop) is substantially more robust to the Stage 1 alone baseline (ST1) on AgeDB [37].

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

Table S4. More results of our 2-stage approach based on ArcFace features (8×8 grid), CosFace features (6× 7), and FaceNet features (3 × 3) across all feature weighting methods which perform slightly better than the Stage 1 alone (ST1) baseline at P@1 when the query is a rotated face (*i.e*. profile faces from CFP [48]).

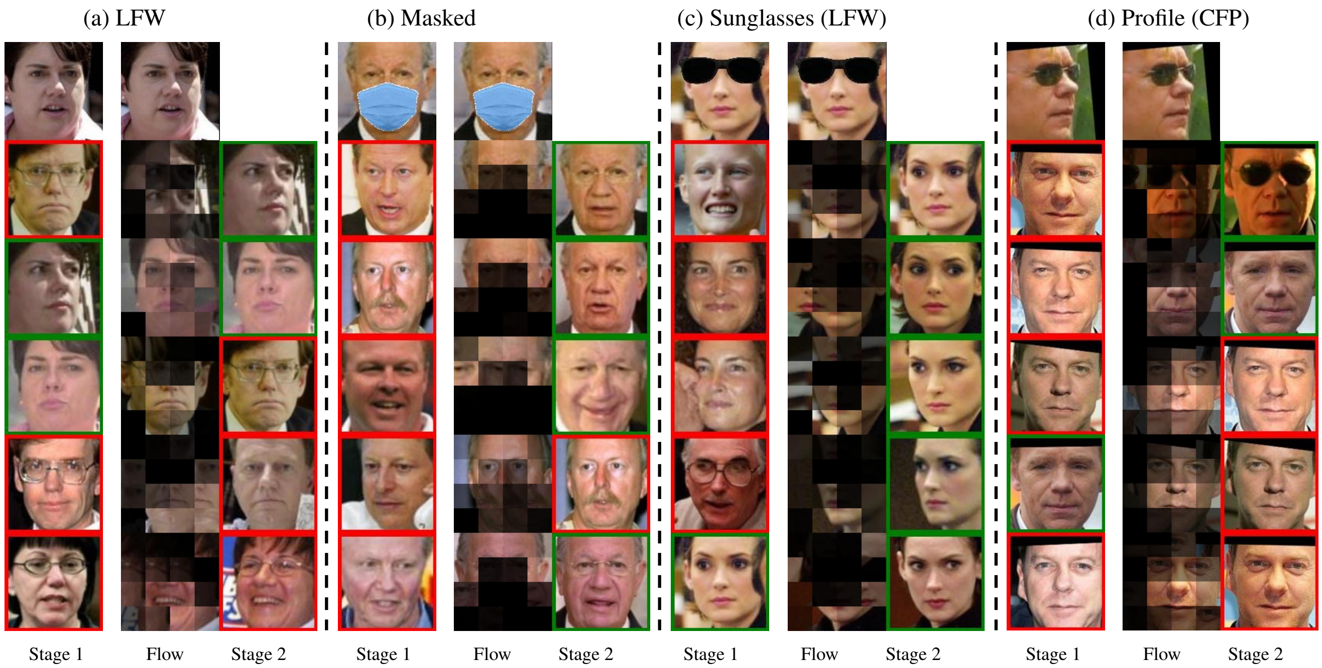


Figure S4. Traditional face identification ranks gallery images based on their cosine distance with the query (top row) at the image-level embedding, which yields large errors upon out-of-distribution changes in the input (*e.g.* masks or sunglasses; b–d). We find that re-ranking the top-k shortlisted faces from Stage 1 (leftmost column) using their patch-wise EMD similarity w.r.t. the query substantially improves the precision (Stage 2) on challenging cases (b–d). The "Flow" visualization (of 4×4) intuitively shows the patch-wise reconstruction of the query face using the most similar patches (*i.e.* highest flow) from the retrieved face.

Dataset	Model	Method	P@1	RP	M@R
TALFW	ArcFace	Cosine	93.49	81.04	80.35
		Uniform	**96.72**	**83.41**	**82.80**
		APC	**96.54**	**82.72**	**82.10**
		SC	**96.71**	**83.39**	**82.78**
	CosFace	Cosine	96.49	83.57	82.99
		SC	**99.14**	**85.03**	**55.27**
		Uniform	**99.14**	**85.56**	**85.11**
		APC	**99.07**	**85.48**	**85.08**
	FaceNet	Cosine	95.33	79.24	78.19
		APC	**97.26**	**80.33**	**79.39**
		Uniform	**97.70**	**80.10**	**79.15**
		SC	**97.59**	**79.85**	**78.89**

Table S5. Our re-ranking 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 and models (ArcFace, CosFace, FaceNet).

Models in MLFW Table 3 [58]	Method	MLFW
Private-Asia, R50, ArcFace	[58]	74.85%
	+ DeepFaceEMD	**76.50%**
CASIA, R50, CosFace	[58]	82.87%
	+ DeepFaceEMD	**87.17%**
MS1MV2, R100, Curricularface	[58]	90.60%
	+ DeepFaceEMD	**91.17%**

Table S6. Using our proposed similarity function consistently improves the face verification results on MLFW (*i.e.* OOD masked images) for models reported in Wang et al. [59]. We use pre-trained models and code by [59].

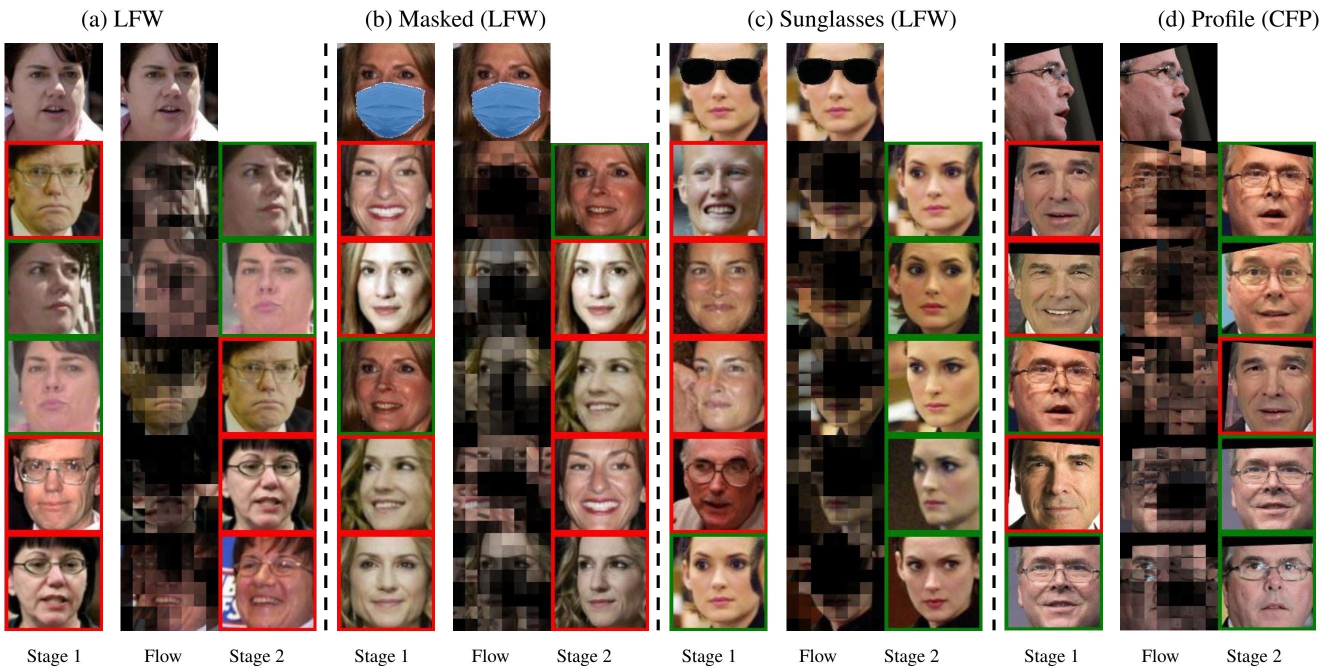


Figure S5. Traditional face identification ranks gallery images based on their cosine distance with the query (top row) at the image-level embedding, which yields large errors upon out-of-distribution changes in the input (*e.g.* masks or sunglasses; b–d). We find that re-ranking the top-k shortlisted faces from Stage 1 (leftmost column) using their patch-wise EMD similarity w.r.t. the query substantially improves the precision (Stage 2) on challenging cases (b–d). The "flow" visualization (of 8×8) intuitively shows the patch-wise reconstruction of the query face using the most similar patches (*i.e.* highest flow) from the retrieved face.

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

Table S7. DeepFace-EMD improved FI on the reported datasets regardless of the pre-processing choice.

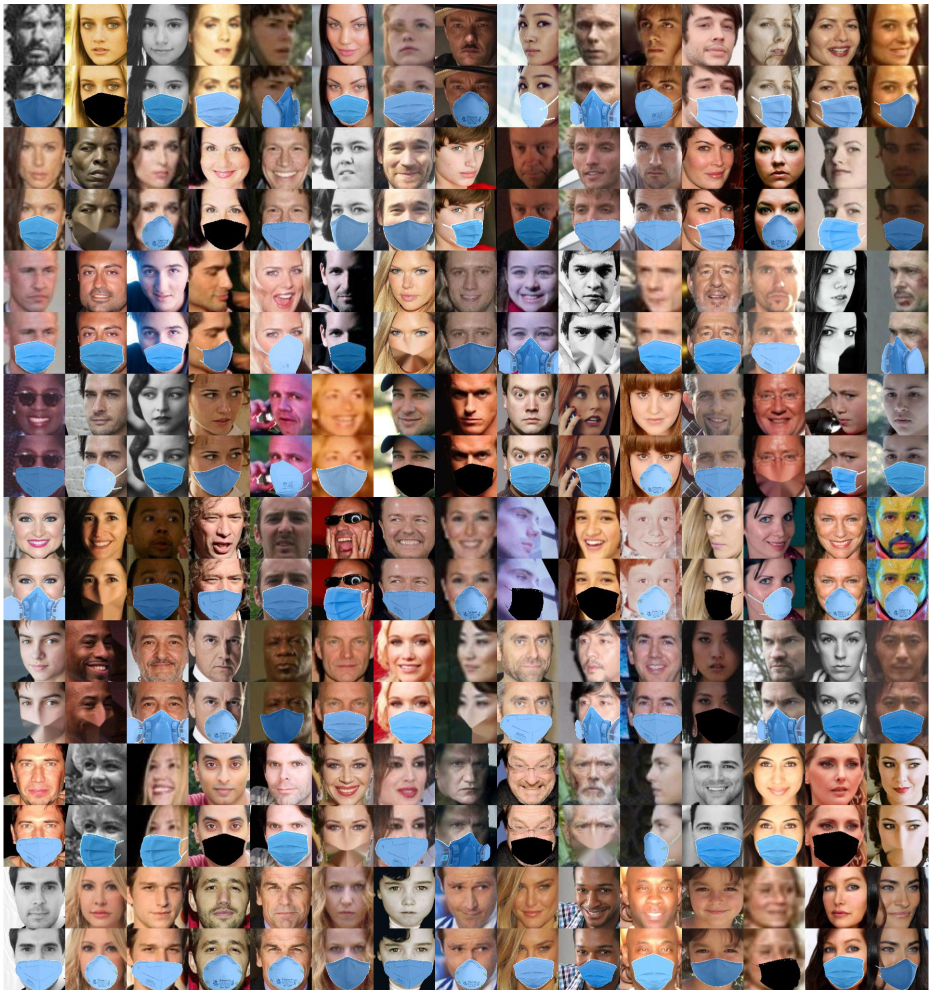


Figure S6. Our CASIA dataset augmented with masked images (generated following the method by [10]) for fine-tuning ArcFace.

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

Table S8. Our 2-stage approach based on ArcFace features (8×8 grid; APC) performs slightly better than the Stage 1 alone (ST1) baseline at P@1 when the query is a rotated face (*i.e*. profile faces from CFP [48]). See Tab. S4 for the results of occlusions on CFP.