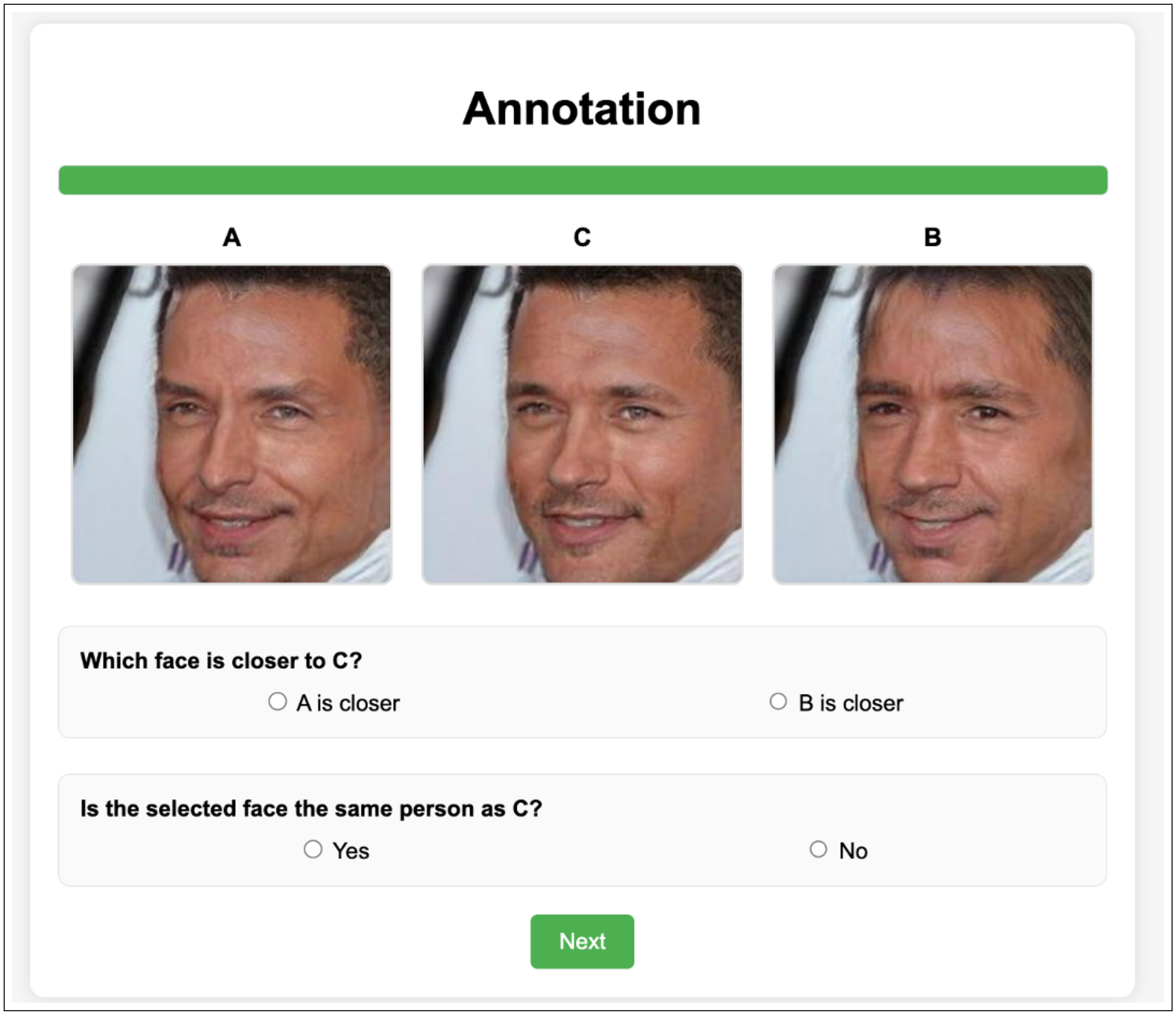
SUPPLEMENTARY MATERIAL

1. DATASET

In this section, supplementary materials related to the dataset are provided. The annotation interface is shown in Figure S1. Table S1 presents the survey results regarding the criteria participants used to select more similar faces after completing the annotation task.



Which face is closer to C?

○ A is closer ○ B is closer

Is the selected face the same person as C?

○ Yes ○ No

Fig. S1. User interface of the annotation tool. In the first question, participants were asked which of the two faces, A or B, is closer to the reference image C. In the second question, they were asked whether the face selected in the first question is the same person as C.

Table S1. Responses to the question, “What aspects did you focus on when judging facial similarity?” ordered by priority, based on a questionnaire conducted after the annotation process.

No.	1st	2nd	3rd	4th and Beyond
1	Eye color	Mouth	Contour	Wrinkles
2	Overall impression	Contour	Expression	Depth of double eyelid
3	Eye shape	Gender	Mouth	Nose
4	Feature arrangement	Age	Gender	Eye color
5	Bone structure	Contour	Nose	Overall impression
6	Overall impression	Eye color	Mouth shape	-
7	Eyes	Eyebrows	Mouth	-
8	Skin	Eyes	Hairstyle	-
9	Eye shape	Eye color	Mouth	Nose
10	Eye color	Eye shape	Nose	Mouth
11	Eyes	Mouth	Wrinkles	-
12	Eye shape	Protrusion depth	Expression	Mouth
13	Eye color	Eye position	Eye shape	Eyebrow shape
14	Eye area	Mouth	Contour	Bone structure
15	Eye area	Contour	Eyebrow shape	Nose shape
16	Eye color	Eye shape	Nose shape	Contour, Eyebrows
17	Eyes	Mouth	Eyebrows	Nose, Hair
18	Eye and nose arrangement	Below the nose	Eye shape	Contour

2. FACE SIMILARITY PREDICTION

In this section, we provide supplementary materials related to face similarity. Figure S2 illustrates examples of successful and failed predictions. Figure S3 shows the similarity distribution between query images and each attribute group. Table S2 presents the results of attribute classification, where the distance D_{I_q,G_i} between a query image and a group is defined as the group mean of distances $d_{I_q,I_{G_i,j}}$ between the query image and images belonging to the group. Table S3 shows the results of attribute classification, where D_{I_q,G_i} is defined as the upper limit of the confidence interval of $d_{I_q,I_{G_i,j}}$. Table S4 presents the results of attribute classification, where D_{I_q,G_i} is defined as the top-k mean of $d_{I_q,I_{G_i,j}}$ within the group. Figure S4 demonstrates the images in each attribute group sorted by similarity to the query image, highlighting the top two most similar and bottom five least similar images.

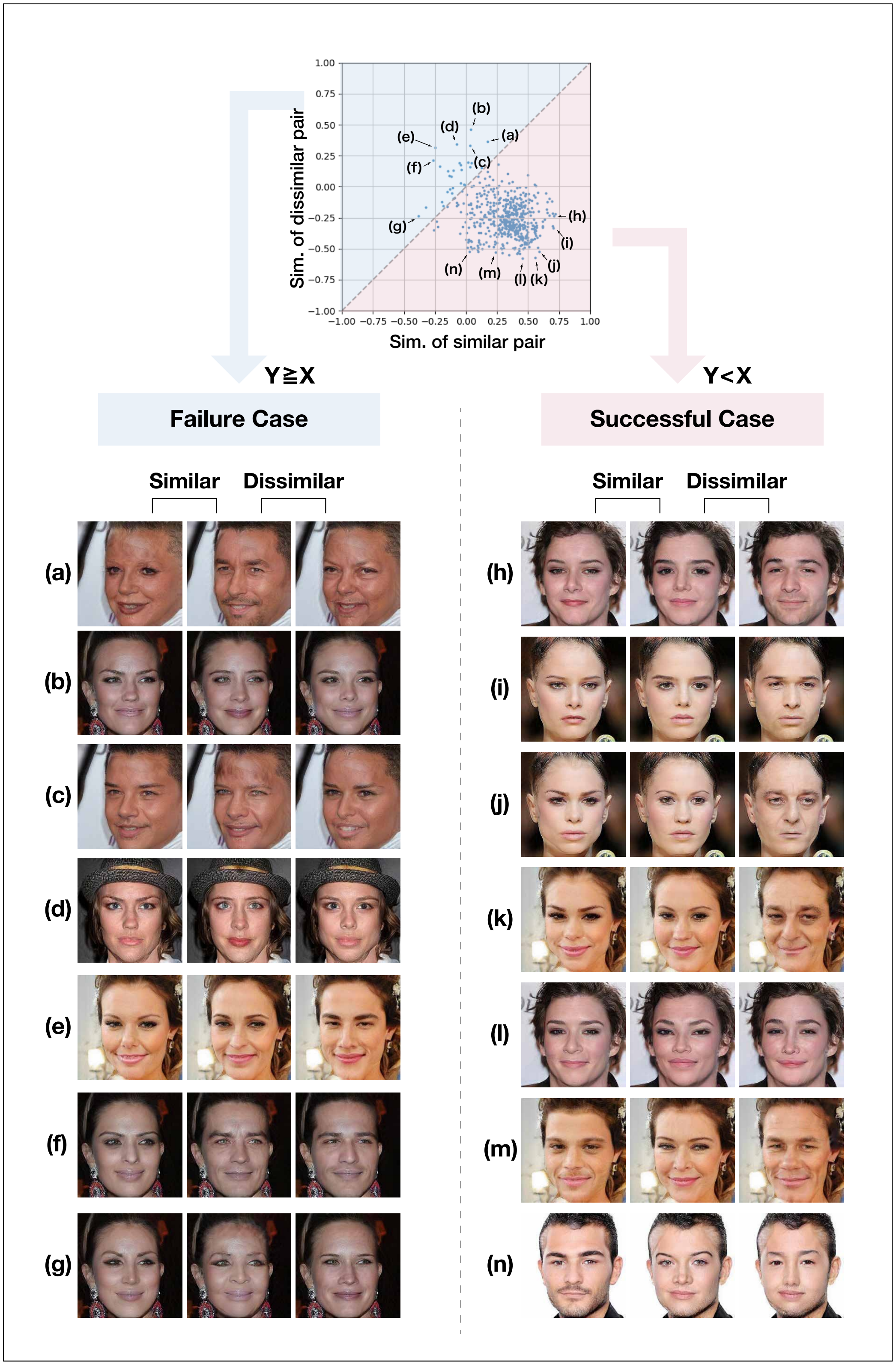


Fig. S2. Examples of successful and failed samples in the similarity prediction task using the evaluation dataset [ii].

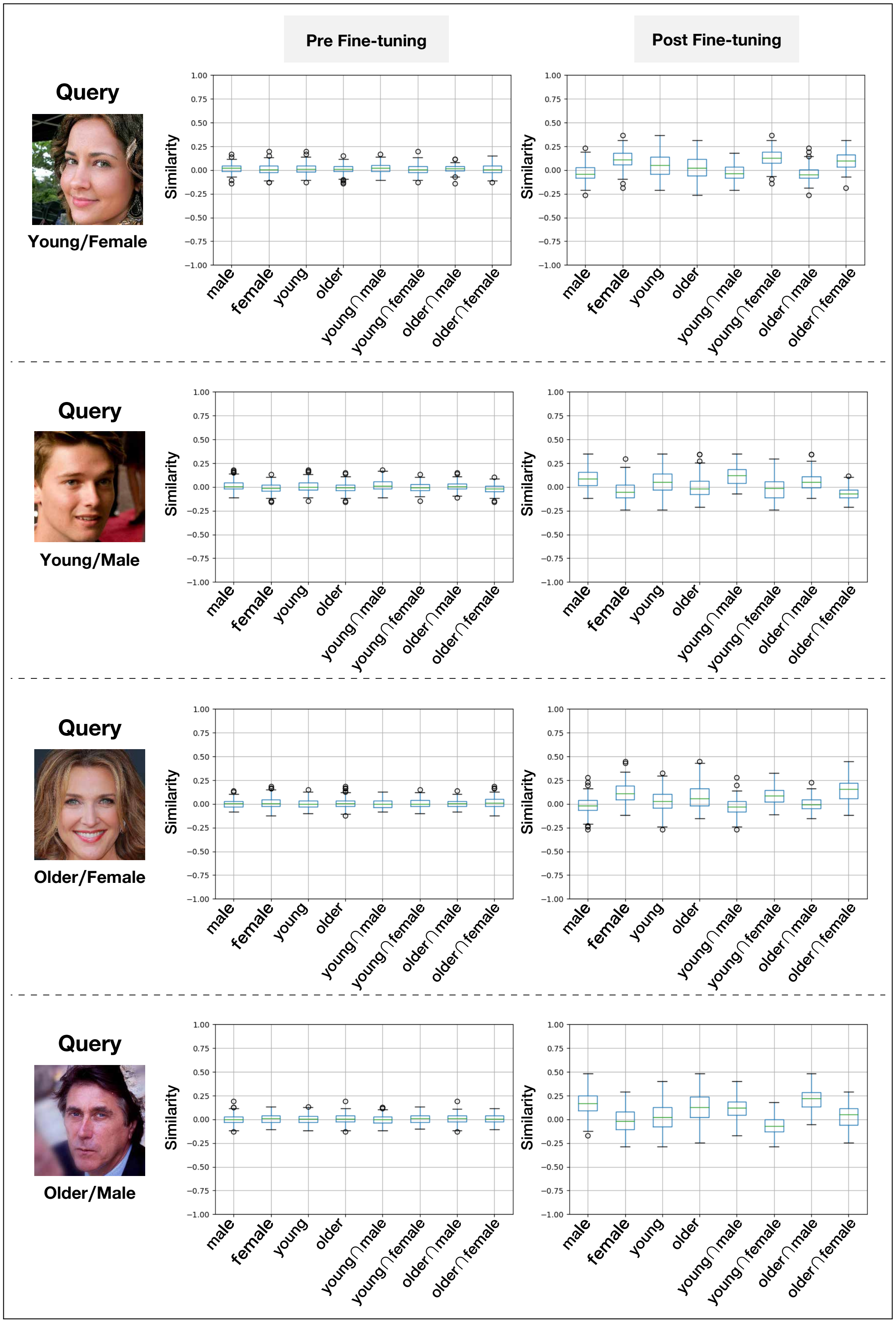


Fig. S3. Distributions of similarity between query images and face swap candidates within each attribute group. The classification for male/female and young/older is binary. When considering multiple attributes simultaneously, classification is performed into four groups: "male and young," "female and young," "male and older," and "female and older."

Table S2. Classification accuracy when D_{I_q,G_i} is the group-wide average of distances $d_{I_q,I_{G_i,j}}$.

	Classification Category	**Precision**	**Recall**	**Accuracy**	**AUC**
Pre fine-tuning	Male	0.680	0.680	0.680	0.680
	Female	0.680	0.680		
	Young	0.596	0.680	0.610	0.610
	Older	0.628	0.540		
	Young∩Male	0.486	0.680	0.740	0.720
	Young∩Female	0.552	0.640	0.730	0.647
	Older∩Male	0.600	0.360	0.780	0.640
	Older∩Female	0.458	0.440	0.730	0.633
Post fine-tuning	Male	0.958	0.920	0.940	0.940
	Female	0.923	0.960		
	Young	0.850	0.680	0.780	0.780
	Older	0.733	0.880		
	Young∩Male	0.826	0.760	0.900	0.853
	Young∩Female	0.720	0.720	0.860	0.813
	Older∩Male	0.800	0.800	0.900	0.867
	Older∩Female	0.667	0.720	0.840	0.800

Table S3. Classification accuracy when D_{I_q,G_i} is the upper bound of the confidence interval of distances $d_{I_q,I_{G_i},j}$.

γ		**Classification Category**	**Precision**	**Recall**	**Accuracy**	**AUC**
0.05	Pre fine-tuning	Male	0.720	0.720	0.720	0.720
		Female	0.720	0.720		
		Young	0.632	0.720	0.650	0.650
		Older	0.674	0.580		
		Young∩Male	0.471	0.640	0.730	0.700
		Young∩Female	0.481	0.520	0.740	0.667
		Older∩Male	0.733	0.440	0.820	0.693
		Older∩Female	0.583	0.560	0.790	0.713
	Post fine-tuning	Male	0.958	0.920	0.940	0.940
		Female	0.923	0.960		
		Young	0.850	0.680	0.780	0.780
		Older	0.733	0.880		
		Young∩Male	0.826	0.760	0.900	0.853
		Young∩Female	0.750	0.720	0.870	0.820
		Older∩Male	0.800	0.800	0.900	0.867
		Older∩Female	0.679	0.760	0.850	0.820
0.01	Pre fine-tuning	Male	0.720	0.720	0.720	0.720
		Female	0.720	0.720		
		Young	0.632	0.720	0.650	0.650
		Older	0.674	0.580		
		Young∩Male	0.471	0.640	0.730	0.700
		Young∩Female	0.481	0.520	0.740	0.667
		Older∩Male	0.733	0.440	0.820	0.693
		Older∩Female	0.583	0.560	0.790	0.713
	Post fine-tuning	Male	0.958	0.920	0.940	0.940
		Female	0.923	0.960		
		Young	0.850	0.680	0.780	0.780
		Older	0.733	0.880		
		Young∩Male	0.826	0.760	0.900	0.853
		Young∩Female	0.720	0.720	0.860	0.813
		Older∩Male	0.800	0.800	0.900	0.867
		Older∩Female	0.667	0.720	0.840	0.800

Table S4: Classification accuracy when D_{I_q,G_i} is the top-k average of distances $d_{I_q,I_{G_i},j}$ within the set G_i.

k		**Category**	**Precision**	**Recall**	**Accuracy**	**AUC**
5	Pre Fine-tuning	Male	0.932	0.820	0.880	0.880
		Female	0.839	0.940		
		Young	0.717	0.760	0.730	0.730
		Older	0.745	0.700		
		Young∩Male	0.630	0.680	0.820	0.773
		Young∩Female	0.656	0.840	0.850	0.847
		Older∩Male	0.750	0.480	0.83	0.713
		Older∩Female	0.720	0.720	0.860	0.813
	Post Fine-tuning	Male	1.000	0.900	0.950	0.950
		Female	0.910	1.000		
		Young	0.759	0.880	0.800	0.800
		Older	0.857	0.720		
		Young∩Male	0.731	0.760	0.870	0.833
		Young∩Female	0.710	0.880	0.880	0.880
		Older∩Male	0.800	0.640	0.870	0.793
		Older∩Female	0.783	0.720	0.880	0.827
10	Pre Fine-tuning	Male	0.953	0.820	0.890	0.890
		Female	0.842	0.960		
		Young	0.678	0.800	0.710	0.710
		Older	0.756	0.620		
		Young∩Male	0.556	0.600	0.780	0.720
		Young∩Female	0.586	0.680	0.800	0.760
		Older∩Male	0.684	0.520	0.820	0.720
		Older∩Female	0.720	0.720	0.860	0.813
	Post Fine-tuning	Male	1.000	0.920	0.960	0.960
		Female	0.926	1.000		
		Young	0.781	0.860	0.810	0.810
		Older	0.844	0.760		
		Young∩Male	0.800	0.800	0.900	0.867
		Young∩Female	0.710	0.880	0.880	0.880
		Older∩Male	0.857	0.720	0.900	0.840
		Older∩Female	0.783	0.720	0.880	0.827
20	Pre Fine-tuning	Male	0.913	0.840	0.880	0.880
		Female	0.852	0.920		
		Young	0.691	0.760	0.710	0.710
		Older	0.733	0.660		
		Young∩Male	0.536	0.600	0.770	0.713
		Young∩Female	0.567	0.680	0.790	0.753
		Older∩Male	0.722	0.520	0.830	0.727
		Older∩Female	0.750	0.720	0.870	0.820
	Post Fine-tuning	Male	1.000	0.920	0.960	0.960
		Female	0.930	1.000		
		Young	0.811	0.860	0.830	0.830
		Older	0.851	0.800		
		Young∩Male	0.833	0.800	0.910	0.873
		Young∩Female	0.733	0.880	0.890	0.887
		Older∩Male	0.864	0.760	0.910	0.860
		Older∩Female	0.792	0.760	0.890	0.847

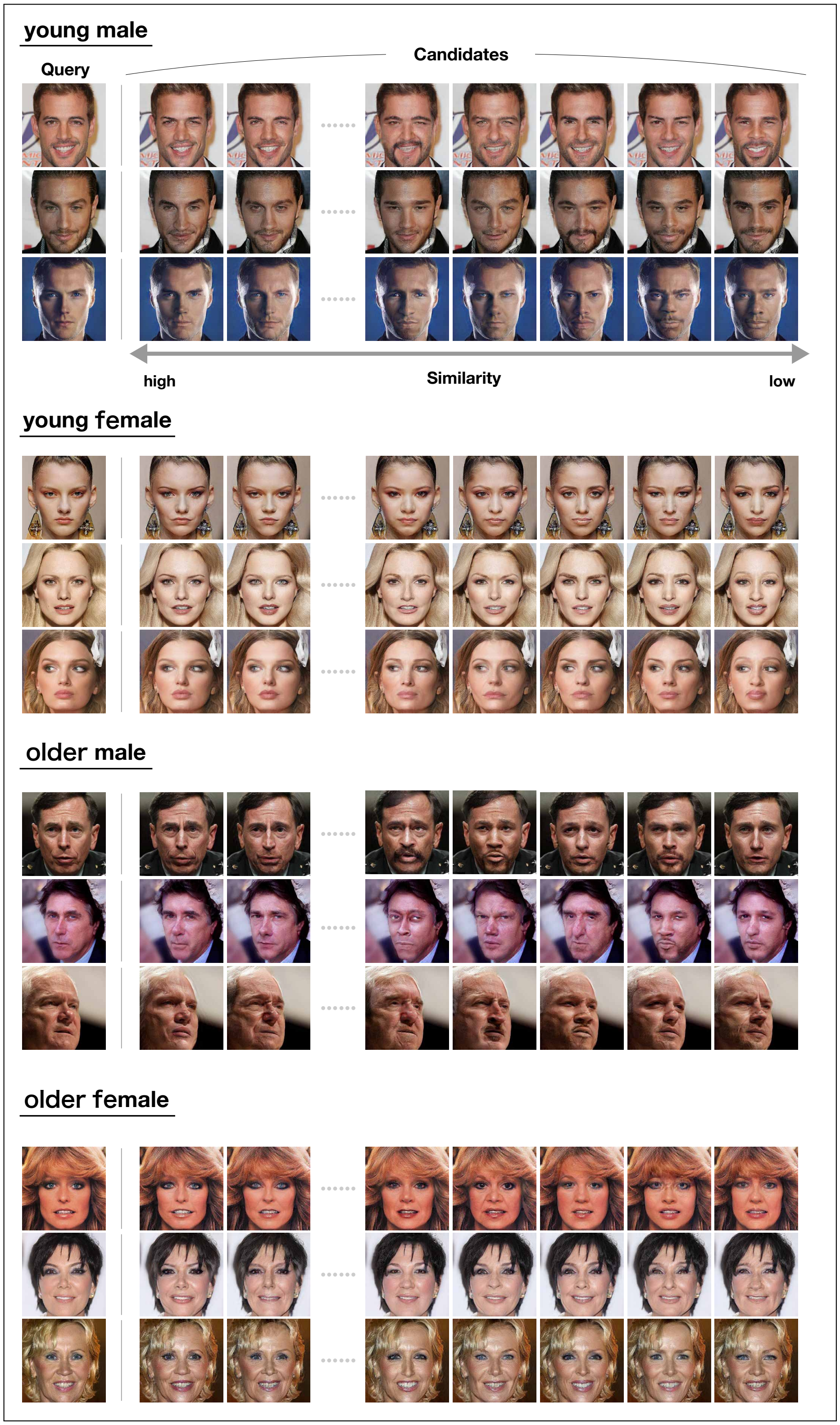


Fig. S4. Based on the proposed similarity metric, face swap candidates were reordered. For query images assigned to the attribute groups "young ∩ male," "young ∩ female," "older ∩ male," and "older ∩ female," candidates within each selected attribute group were sorted by similarity. The top two most similar and bottom five least similar candidates were displayed for each group.