2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

On Face Segmentation, Face Swapping, and Face Perception

Yuval Nirkin[1], Iacopo Masi[2], Anh Tuấn Trần[2], Tal Hassner[1,3], and Gérard Medioni[2]
[1] *The Open University of Israel, Israel*
[2] *Institute for Robotics and Intelligent Systems, USC, CA, USA*
[3] *Information Sciences Institute, USC, CA, USA*

Figure 1: *Inter-subject swapping.* LFW G.W. Bush photos swapped using our method onto very different subjects and images. Unlike previous work [4], [17], we do not select convenient targets for swapping.

***Abstract*—We show that even when face images are unconstrained and arbitrarily paired, face swapping between them is quite simple. To this end, we make the following contributions. (a) Instead of tailoring systems for face segmentation, as others previously proposed, we show that a standard fully convolutional network (FCN) can achieve remarkably fast and accurate segmentations, provided that it is trained on a rich enough example set. For this purpose, we describe novel data collection and generation routines which provide challenging segmented face examples. (b) We use our segmentations for robust face swapping under unprecedented conditions. (c) Unlike previous work, our swapping is robust enough to allow for extensive quantitative tests. To this end, we use the Labeled Faces in the Wild (LFW) benchmark and measure the effect of intra- and inter-subject face swapping on recognition. We show that our intra-subject swapped faces remain as recognizable as their sources, testifying to the effectiveness of our method. In line with established perceptual studies, we show that better face swapping produces less recognizable inter-subject results (see, e.g., Fig. 1). This is the first time this effect was quantitatively demonstrated by machine vision systems.**

***Keywords*-Face swapping, segmentation, recognition**

I. Introduction

Swapping faces means transferring a face from a *source* photo onto a face appearing in a *target* photo, attempting to generate realistic, unedited looking results. Although face swapping today is often associated with viral Internet memes [31], it is actually far more important than this practice may suggest: Face swapping can also be used for preserving privacy [4], [30], digital forensics [31] and as a potential face specific data augmentation method [29] especially in applications where training data is scarce (e.g., facial emotion recognition [24]).

Going beyond particular applications, face swapping is also an excellent opportunity to develop and test essential face processing capabilities: When faces are swapped between arbitrarily selected, unconstrained images, there is no guarantee on the similarity of viewpoints, expressions, 3D face shapes, genders or any other attribute that makes swapping easy [17]. In such cases, swapping requires robust and effective methods for face alignment, segmentation, 3D shape estimation (though we will later challenge this assertion), expression estimation and more.

We describe a face swapping method and test it in settings where no control is assumed over the images *or their pairings*. We evaluate our method using extensive quantitative tests at a scale never before attempted by other face swapping methods. These tests allow us to measure the effect face swapping has on machine face recognition, providing insights from the perspectives of both security applications and face perception.

Technically, we focus on face segmentation and the design of a face swapping pipeline. Our contributions include:

- *Semi-supervised labeling of face segmentation.* We provide a novel means of generating a rich image set with face segmentation labels, by using using motion cues and 3D data augmentation. The data we collect is used to train a FCN to segment faces faster and more accurately than existing methods.
- *Face swapping pipeline.* We describe a face swapping pipeline and show it to work well on images and image pairs of unprecedented difficulty.
- *Quantitative tests.* Despite over a decade of work and contrary to other face processing tasks (e.g., recognition), face swapping methods were never quantitatively tested. We offer the first quantitative test protocols for intra- and inter-subject face swapping systems.

Our qualitative results show that our swapped faces are as compelling as those produced by others, if not more. Our quantitative tests further show that our intra-subject face

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00024

swapping has little effect on face verification accuracy: our swapping does not introduce artifacts or otherwise changes these images in ways which affect subject identities.

We report inter-subject results on randomly selected pairs. These tests require facial appearance to change, sometimes substantially, in order to naturally blend source faces into their new surroundings. We show that this changes them, making them less recognizable. Though this perceptual phenomenon was described over two decades ago by Sinha and Poggio [35] in their well-known Clinton-Gore illusion, we are unaware of previous quantitative reports on how this applies to machine face recognition.

For code and deep models, please see our project page.[1]

II. Related Work

Face segmentation. To swap only faces, without their surrounding context or occlusions, we require per-pixel segmentation labels. Previous methods designed for this purpose include the work of [28] which segment individual facial regions (e.g., eyes, mouth) but not the entire face. An example based method was proposed in [36]. More recently, [9] segmented faces by alternating between segmentation and landmark localization using deformable part models. They report state of the art performance on the Caltech Occluded Faces in the Wild (COFW) dataset [6].

Two recent methods proposed to segment faces using deep neural networks. In [26] a network was trained to simultaneously segment multiple facial regions, including the entire face. This method was used in the face swapping method of [17], but can be slow. The very recent method of [34] recently outperformed [9] on COFW as well as reported real-time processing speeds by using a deconvolutional neural network.

Face swapping. Methods for swapping faces were proposed as far back as 2004 [5] with fully automatic techniques described nearly a decade ago [4]. These methods were originally offered in response to privacy preservation concerns: Face swapping can be used to obfuscate identities of subjects appearing in publicly available photos, as a substitute to face pixelation or blurring [4], [5], [30]. Since then, however, many of their applications seem to come from recreation [17] or entertainment (e.g., [1], [38]).

Regardless of the application, previous face swapping systems often share several key aspects. First, some methods restrict the target photos used for transfer. Given an input source face, they search through large face albums to choose ones that are easy targets for face swapping [4], [8], [17]. Such targets are those which share similar appearance properties with the source, including facial tone, pose, expression and more. Though our method can be applied in similar settings, our tests focus on more extreme conditions, where the source and target images are arbitrarily selected and can be (often are) substantially different.

Second, most previous methods estimate the structure of the face. Some methods estimate 3D facial shapes [1], [4], [25], by fitting 3D Morphable Face Models (3DMM). Others instead estimate dense 2D active appearance models [40]. This is presumably done in order to correctly map textures across different individual facial shapes.

Finally, deep learning was used to transfer faces [19], as if they were styles transfered between images. This method, however, requires the network to be trained for each source image and thus can be impractical in many applications.

III. Swapping faces in unconstrained images

Fig. 3 summarizes our face swapping method. When swapping a face from a source image, $\mathbf{I}_S$, to a target image, $\mathbf{I}_T$, we treat both images the same, apart from the final stage (Fig. 3(d)). Our method first localizes 2D facial landmarks in each image (Fig. 3(b)). We use an off-the-shelf detector for this purpose [16]. Using these landmarks, we compute 3D pose (viewpoint) and modify the 3D shape to account for expression. These steps are discussed in Sec. III-A.

We next segment faces from backgrounds and occlusions (Fig. 3(c)) using a FCN trained to predict per-pixel face visibility (Sec. III-B). We describe how we generate rich labeled data to train our FCN. Finally, the source is efficiently warped onto the target using the two aligned 3D face shapes as proxies and blended onto the target (Sec. III-C).

A. Fitting 3D face shapes

To enrich our set of examples for training the segmentation network (Sec. III-B) we explicitly model 3D face shapes. These 3D shapes are also used as proxies to transfer textures from one face onto another, when swapping faces (Sec. III-C). We experimented with two alternative methods of obtaining these 3D shapes.

The first, inspired by [12] uses a generic 3D face, making no attempt to fit its shape to the face in the image aside from pose (viewpoint) alignment. We, however, also estimate facial expressions and modify the 3D face accordingly.

A second approach uses the recent state of the art, deep method for single image 3D face reconstruction [37]. It was shown to work well on unconstrained photos such as those considered here. To our knowledge, this is the only method quantitatively shown to produce invariant, discriminative and accurate 3D shape estimations. The code they released regresses 3D Morphable face Models (3DMM) in neutral pose and expression. We extend it by aligning 3D shapes with input photos and modifying the 3D faces to account for facial expressions.

3D shape representation and estimation. Whether generic or regressed, we use the popular Basel Face Model (BFM) [32] to represent faces and the 3DDFA Morphable

[1] www.openu.ac.il/home/hassner/projects/faceswap

(a) Input (b) Fit 3D Shape, pose and expr. (c) Segment (d) Transfer (e) Output

Figure 3: *Method overview.* (a) Source (top) and target (bottom) input images. (b) Detected facial landmarks used to establish 3D pose and facial expression for a 3D face shape (Sec. III-A). (c) Our segmentation of Sec. III-B (red) overlaid on the projected 3D face (gray). (d) Source transfered onto target without blending, and the final results (e) after blending (Sec. III-C).

Model [41] for expressions. These are both publicly available 3DMM representations. More specifically, a 3-D face shape $\mathbf{V} \subset \mathbb{R}^3$ is modeled by combining the following independent generative models:

$$\mathbf{V} = \widehat{\mathbf{v}} + \mathbf{W}_S\ \boldsymbol{\alpha} + \mathbf{W}_E\ \boldsymbol{\gamma}. \tag{1}$$

Here, vector $\widehat{\mathbf{v}}$ is the mean face shape, computed over aligned facial 3D scans in the Basel Faces collection and represented by the concatenated 3D coordinates of their 3D points. When using a generic face shape, we use this average face. Matrices $\mathbf{W}_S$ (shape) and $\mathbf{W}_E$ (expression) are principle components obtained from the 3D face scans. Finally, $\boldsymbol{\alpha}$ is a subject-specific 99D parameter vector estimated separately for each image and $\boldsymbol{\gamma}$ is a 29D parameter vector for expressions. To fit 3D shapes and expressions to an input image, we estimate these parameters along with camera matrices.

To estimate per-subject 3D face shapes, we regress $\boldsymbol{\alpha}$ using the deep network of [37]. They jointly estimate 198D parameters for face shape and texture. Dropping the texture components, we obtain $\boldsymbol{\alpha}$ and back-project the regressed face by $\widehat{\mathbf{v}} + \mathbf{W}_S\ \boldsymbol{\alpha}$, to get the estimated shape in 3D space.

Pose and expression fitting. Given a 3D face shape (generic or regressed) we recover its pose and adjust its expression to match the face in the input image. We use the detected facial landmarks, $\mathbf{p} = \{\mathbf{p}_i\} \subset \mathbb{R}^2$, for both purposes. Specifically, we begin by solving for the pose, ignoring expression. We approximate the positions in 3D of the detected 2D facial landmarks $\tilde{\mathbf{V}} = \{\tilde{\mathbf{V}}_\mathbf{i}\}$ by:

$$\tilde{\mathbf{V}} \approx f(\widehat{\mathbf{v}}) + f(\mathbf{W}_S)\ \boldsymbol{\alpha}, \tag{2}$$

where $f(\cdot)$ is a function selecting the landmark vertices on the 3D model. The vertices of all BFM faces are registered so that the same vertex index corresponds to the same facial feature in all faces. Hence, f need only be manually specified once, at preprocessing. From f we get 2D-3D correspondences, $\mathbf{p}_i \leftrightarrow \tilde{\mathbf{V}}_i$, between detected facial features and their corresponding points on the 3D shape. Similarly to [11], we use these correspondences to estimate 3D pose, computing 3D face rotation, $\mathbf{R} \in \mathbb{R}^3$, and translation vector $\mathbf{t} \in \mathbb{R}^3$ using the EPnP solver [23].

Following pose estimation, we estimate the expression parameters in vector $\boldsymbol{\gamma}$ by formulating expression estimation as a bounded linear problem:

$$\delta_R\Big(P(\mathbf{R},\mathbf{t})\big(f(\widehat{\mathbf{v}}) + f(\mathbf{W}_S)\ \boldsymbol{\alpha} + f(\mathbf{W}_E)\ \boldsymbol{\gamma}\big)\Big) = \delta_R(\mathbf{p}),$$
$$\text{with } \big|\boldsymbol{\gamma}_j\big| \leq 3\ \sigma_j \qquad \forall\ j = \{1 \ldots 29\} \tag{3}$$

where $\delta_R(\cdot)$ is a visibility check that removes occluded points given the head rotation $\mathbf{R}$; $P(\mathbf{R},\mathbf{t})$ is the projection matrix, given the extrinsic parameters ($\mathbf{R}$,$\mathbf{t}$); and σ_j is the standard deviation of the j-th expression component in $\boldsymbol{\gamma}$. This problem can be solved using any constrained linear least-squares solver.

B. Deep face segmentation

Our method uses a FCN to segment the visible parts of faces from their context and occlusions. Other methods previously tailored novel network architectures for this task (e.g., [34]). We show that excellent segmentation results can be obtained with a standard FCN, provided that it is trained on plenty of rich and varied examples.

Obtaining enough diverse images with ground truth segmentation labels can be hard: Saito et al. [34], for example, used manually segmented LFW faces and a semi-automatic segmentation method [7] for this purpose. These labels were costly to produce and limited in their variability and number. We instead propose a novel means of generating numerous training examples with little manual effort and show that a *standard FCN* trained on these examples outperforms state of the art face segmentation results.

Semi-supervised training data collection. We produce large quantities of segmentation labeled face images by using *motion cues* in unconstrained face videos. To this

end, we process videos from the recent IARPA Janus CS2 dataset [18]. These videos portray faces of different poses, ethnicities and ages, viewed under widely varying conditions. We used 1,275 videos of subjects not included in LFW, of the 2,042 CS2 videos (309 subjects out of 500).

Given a video, we produce a rough, initial segmentation using a method based on [10]. Specifically, we keep a hierarchy of regions with stable region boundaries computed with dense optical flow. Though these regions may be over- or under-segmented, they are computed with temporal coherence and so these segments are consistent across frames.

We use the method of [16] to detect faces and facial landmarks in each of the frames. Facial landmarks were then used to extract the face contour and extend it to include the forehead. All the segmented regions generated above, that did not overlap with a face contour are then discarded. All intersecting segmented regions are further processed using a simple interface which allows browsing the entire video, selecting the partial segments of [10] and adding or removing them from the face segmentation using simple mouse clicks. Fig. 4(a) shows the interface used in the semi-supervised labeling. A selected frame is typically processed in about five seconds. In total, we used this method to produce 9,818 segmented faces, choosing anywhere between one to five frames per video in a little over a day of work.

Occlusion augmentation. This collection is further enriched by adding synthetic occlusions. To this end, we explicitly use 3D information estimated for our example faces. Specifically, we estimate 3D face shape for our segmented faces, using the method described in Sec. III-A. We then use computer graphic (CG) 3D models of various objects (e.g., sunglasses) to modify the faces. We project these CG models onto the image and record their image locations as synthetic occlusions. Each CG object added 9,500 face examples. The detector used in our system [16] failed to accurately localize facial features on the remaining 318 faces, and so this augmentation was not applied to them.

Finally, an additional source of synthetic occlusions was supplied following [34] by overlaying hand images at various positions on our example images. Hand images were taken from the egohands dataset of [3]. Fig 4(b) shows a synthetic hand augmentation and Fig 4(c) a sunglasses augmentation, along with their resulting segmentation labels.

C. Face swapping and blending

Face swapping from a source $\mathbf{I}_S$ to target $\mathbf{I}_T$ proceeds as follows. The 3D shape associated with the source, $\mathbf{V}_S$, is projected down onto $\mathbf{I}_S$ using its estimated pose, $P(\mathbf{R}_S, \mathbf{t}_S)$ (Sec. III-A). We then sample the source image using bilinear interpolation, to assign 3D vertices projected onto the segmented face (Sec. III-B) with intensities sampled from the image at their projected coordinates.

The shapes for both source and target, $\mathbf{V}_S$ and $\mathbf{V}_T$ correspond in the indices of their vertices. We can therefore directly transfer these sampled intensities from all vertices $\mathbf{v}_i \in \mathbf{V}_S$ to $\mathbf{v}_i \in \mathbf{V}_T$. This provides texture for the vertices corresponding to visible regions in $\mathbf{I}_S$ on the target 3D shape. We now render $\mathbf{V}_T$ onto $\mathbf{I}_T$, using the estimated target pose $(\mathbf{R}_T, \mathbf{t}_T)$, masking the rendered intensities using the target face segmentation (see Fig. 3(d)). Finally, the rendered, source face is blended-in with the target context using an off the shelf method [33].

Figure 4: (a) Interface used for semi-supervised labeling. (b-c) Augmented examples and segmentation labels for occlusions due to (b) hands and (c) synthetic sunglasses.

Method	mean IOU	Global	ave(face)	FPS
Struct. Forest [15]*	–	83.9	88.6	–
RPP [39]	72.4	–	–	0.03
SAPM [9]	83.5	88.6	87.1	–
Liu *et al.* [26]	72.9	79.8	89.9	0.29
Saito *et al.* [34]* *+GraphCut*	**83.9**	88.7	92.7	43.2
Us (no occlusion augmentation at training)	81.6	87.4	93.3	**48.6**
Us	83.7	**88.8**	**94.1**	**48.6**

Table I: *COFW [6] segmentation results.*

IV. Experiments

We performed comprehensive experiments in order to test our method, both qualitatively and quantitatively. Runtimes were all measured on an Intel Core i7 4820K computer with 32GB DDR4 RAM and an NVIDIA GeForce Titan X. Using the GPU, our system swaps faces at 1.3 fps. On the CPU, this is slightly slower, performing at 0.8 fps.

A. Face segmentation results

Qualitative face segmentation results are provided in Fig. 3 and 5, visualized following [34] to show segmented regions (red) overlaying the aligned 3D face shapes, projected onto the faces (gray).

We provide also quantitative tests, comparing the accuracy of our segmentations to existing methods. We follow the evaluation procedure described by [9], testing the 507 face photos in the COFW dataset [6]. Previous methods included the regional predictive power (RPP) estimation [39], Structured Forest [15], segmentation-aware part model (SAPM) [9], the deep method of [26], and [34]. WE provide results also for our method, trained without out occlusion augmentation (Sec. III-B). Note that Structured

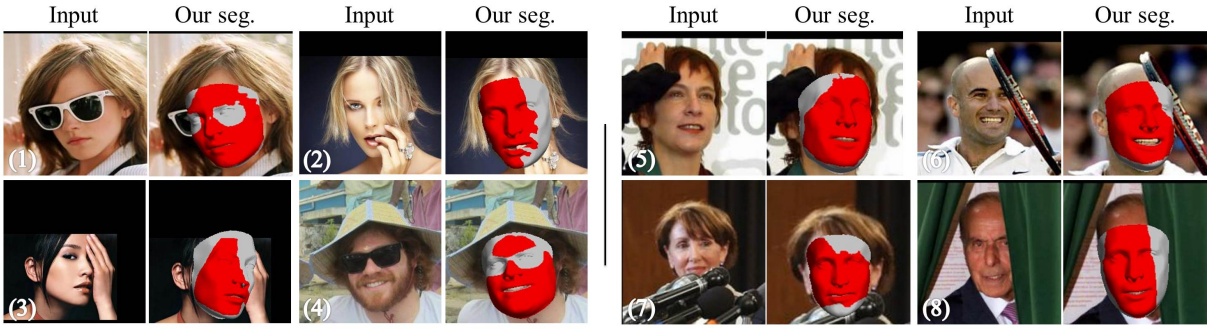

Figure 5: *Qualitative segmentation results from the COFW (1-4) and LFW (5-8) data sets.*

Figure 6: *Qualitative LFW inter-subject face swapping results.* Examples were selected to represent extremely different *poses* (4), *genders* (1,2), *expressions* (1), *ethnicities* (1,3), *ages* (3,4) and *occlusions* (1).

Forest [15] and [34] used respectively 300 and 437 images for testing, without reporting which images were used. Result for [26] was computed by us, using their code, out of the box, but optimizing for the segmentation threshold which provided the best accuracy.

Accuracy is measured using the standard *intersection over union* (IOU) metric, comparing predicted segmentations with manually annotated ground truth masks from [15], as well as two metrics from [15]: *global* – overall percent of correctly labeled pixels – and *ave(face)*, the average face pixel recall. Tab. I reports these results along with run times. Our method is the fastest yet achieves comparable result with the state of the art. Note that we use the same GPU model as [34] and measure run time for [26] ourselves.

B. Qualitative face-swapping results

We provide face swapping examples produced on unconstrained LFW images [14] using randomly selected targets in Fig. 1, 3, and 6. We chose these examples to demonstrate a variety of challenging settings. In particular, these results used source and target faces of widely different poses, occlusions and facial expressions. To our knowledge, previous work never showed results for such challenging settings.

In addition, Fig. 7 shows a qualitative comparison with the very recent method [17] using the same source-target pairs. We note that [17] used the segmentation of [26] which we show in Sec. IV-A to perform worst than our own. This is qualitatively evident in Fig. 7 by the face hairlines. Fig. 7 also provides results from the publicly available code of Kowalski [20] and Hrastnik [13]. In both cases, absence of a segmentation is clearly evident.

C. Quantitative tests

Similarly to previous work, we offer qualitative results to visualize the quality of our swapped faces (Sec. IV-B). Unlike others, however, we also offer extensive quantitative tests designed to measure the effect of swapping on the perceived identity of swapped faces. To this end we propose two test protocols, motivated by the following assumptions.

Assumption 1. Swapping faces between images of different subjects (i.e., *inter-subject swapping*) changes facial *context* (e.g., hair, skin tone, head shape). Effective swapping must therefore modify source faces, sometimes substantially, to blend them naturally into their new contexts thereby producing faces that look less like the source subjects.

Assumption 2. If a face is swapped from source to target photos of the same person (*intra-subject swapping*), the

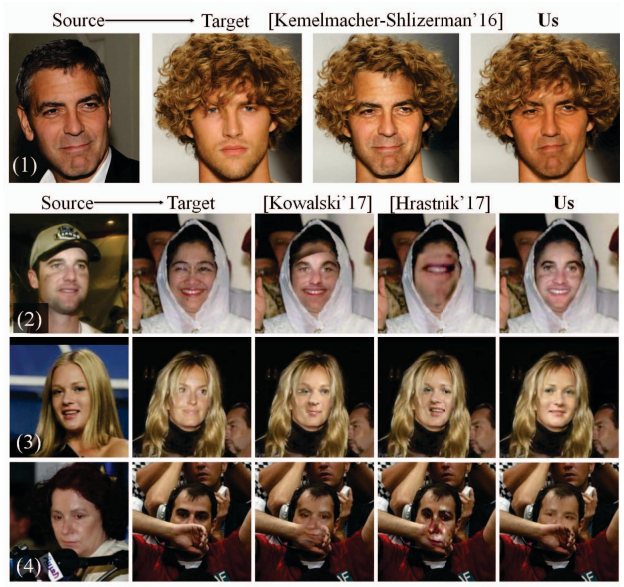

Figure 7: *Comparison with previous face swap methods* (1) Result published by Kemelmacher-Shlizerman [17]. (2-4) Results obtained by the public implementations of Kowalski [20] and Hrastnik [13].

output of an effective swapping method should easily be recognizable as the person in the source photo as the two photos share the same context.

The first assumption is based on well-known trait of human visual perception: Face recognition requires both internal and external cues (faces and their context) to recognize faces. This idea was claimed by the seminal work of [35] and extensively studied in the context of biological visual systems (e.g., [2]). To our knowledge, it was never explored for machine recognition systems and never quantitatively. The robustness of our method allows us to do just that.

The second assumption is intended to verify that when the context remains the same (the same subject) swapping does not change facial appearances in a way which makes faces less recognizable. This ensures that the swapping method does not introduce artifacts or changes facial appearances.

To test these assumptions, we produce modified (face swapped) versions of the LFW benchmark [14]. We estimate how recognizable faces appear after swapping by using a publicly available, state of the art face recognition system in lieu of a large scale human study. Though the recognition abilities of humans and machines may be different, modern systems already claim human or even super-human accuracy [27]. We therefore see the use of a state of the art machine system as an adequate surrogate to human studies which often involve problems of their own [22].

Face verification. We use the ResFace101 [29] face recognition system to test if faces remain recognizable after swapping. ResFace101 obtained near perfect verification results on LFW, yet it was not optimized for that benchmark and tested also on IJB-A [18]. Moreover, it was trained on synthetic face views not unlike the ones produced by face swapping. For these reasons, we expect ResFace101 to be well suited for our purposes. Recognition is measured by 100%-EER (Equal Error Rate), accuracy (Acc.), and normalized Area Under the Curve (nAUC). Finally, we provide ROC curves for all our tests.

Inter-subject swapping verification protocols. We begin by measuring the effect of inter-subject face swapping on face verification accuracy. To this end, we process all faces in the LFW benchmark, swapping them onto photos of *other, randomly selected subjects*. We make no effort to verify the quality of the swapped results and if swapping failed, we treat the result as any other image.

We use the original LFW test protocol with its same/not-same subject pairs. Our images, however, present the original faces with possibly very different contexts. Specifically, let $(\mathbf{I}_i^1, \mathbf{I}_i^2)$ be the i-th LFW test image pair. We produce $\widehat{\mathbf{I}}_i^1$, the swapped version of $\mathbf{I}_i^1$, by randomly picking another LFW subject and image from that subject as a target, taking $\mathbf{I}_i^1$ as the source. We then do the same for $\mathbf{I}_i^2$ to obtain $\widehat{\mathbf{I}}_i^2$.

Matching pairs of swapped images, however, can obscure changes to both images which make the source faces equally unrecognizable: Such tests only reflect the similarity of swapped images to each other, not to their sources. We therefore test verification on benchmark pairs comparing original vs. swapped images. This is done twice, once on pairs $(\widehat{\mathbf{I}}_i^1, \mathbf{I}_i^2)$, the other on pairs $(\mathbf{I}_i^1, \widehat{\mathbf{I}}_i^2)$. We then report the average results for both trials. We refer to these tests as *face preserving* tests.

We also performed *context preserving* tests: These use benchmark image pairs as *targets* rather than sources. Thus, they preserve the context of the original LFW images, not the faces. By doing so, we can measure the effect of context on recognition. This test setup is reminiscent of the *inverse mask* tests performed by [21]. Their tests were designed to measure how well humans recognize LFW faces if the face was cropped out without being replaced, and showed this led to a drop in recognition. Unlike them, our images contain faces of other subjects swapped in place of the original faces, and so are more realistic.

Inter-subject swapping results. We provide verification results for both face preserving and context preserving inter-subject face swapping in Tab. II and ROC curves for the various tests in Fig. 8. Our results include ablation studies, showing accuracy with a generic face and no segmentation (*Generic*), with an estimated 3D face shape (Sec. III-A) and no segmentation (*Est. 3D*), with a generic face and segmentation (*Seg.*) and with an estimated 3D shape and face segmentation (*Est. 3D+Seg.*).

The face preserving results in Tab. II (bottom) are con-

Method	100%-EER	Acc.	nAUC
Baseline (ResFace101)	*98.10±0.90*	*98.12±0.80*	*99.71±0.24*
Context preserving (face swapped out)			
Generic	64.58±2.10	64.56±2.22	69.94±2.24
Est. 3D	69.00±1.43	68.93±1.19	75.58±2.20
Seg.	68.93±1.98	69.00±1.93	76.06± 2.15
Est. 3D+Seg.	73.17±1.59	72.94±1.39	80.77±2.22
Face preserving (face swapped in)			
Generic	92.28±1.37	92.25±1.45	97.55±0.71
Est. 3D	88.77±1.50	88.53±1.25	95.53±0.99
Seg.	89.92±1.48	89.98±1.36	96.17±0.93
Est. 3D+Seg.	86.48±1.74	86.38±1.50	93.71±1.42

Table II: *Inter-subject face swapping.* Ablation study.

Method	100%-EER	Acc.	nAUC
Baseline (VGGFace)	*97.23±0.88*	*97.35±0.77*	*99.54±0.30*
Baseline (ResFace101)	*98.10±0.90*	*98.12±0.80*	*99.71±0.24*
Generic	97.02±0.98	97.02±0.97	99.53±0.31
Est. 3D	97.05±0.98	97.03±1.01	99.52±0.32
Seg.	97.12±1.09	97.08±1.07	99.53±0.31
Est. 3D+Seg.	97.12±1.09	97.12±0.99	99.52±0.31
Est. 3D+Seg.	96.65±0.85	96.63±0.92	99.45±0.29

Table III: *Intra-subject face swapping.* Ablation study.

sistent with our *Assumption 1*: The more the source face is modified, by estimating 3D shape and better segmenting the face, the less it is recognizable as the original subject and the lower the verification results. Using a simple generic shape and no segmentation provides ∼8% *better* accuracy than using our the entire pipeline. Importantly, just by estimating 3D face shapes, accuracy drops by ∼3.5% compared to using a simple generic face shape.

Unsurprisingly, the context preserving results in Tab. II (top) are substantially lower than the face preserving tests. Unlike the face preserving tests, however, the harder we work to blend the randomly selected source faces into their contexts, the better recognition becomes. By better blending the sources into the context, more of the context is retained and the easier it is to verify the two images based on their context without the face itself misleading the match.

Intra-subject swapping verification protocols and results. To test our second assumption, we again process the LFW benchmark, this time swapping faces between different images of the *same subjects* (*intra-subject* face swapping). Of course, all *same* labeled test pairs, by definition, belong to subjects that have at least two images, and so this did not affect these pairs. *Not-same* pairs, however, sometimes include images from subjects which have only a single image. To address this, we replaced them with others for which more than one photo exists.

We again run our entire evaluation twice: once, swapping the first image in each test pairs keeping the second unchanged, and vice versa. Our results average these two trials. Results obtained using different components of our system are provided in Tab. III and Fig. 8. These show that even under extremely different viewing conditions,

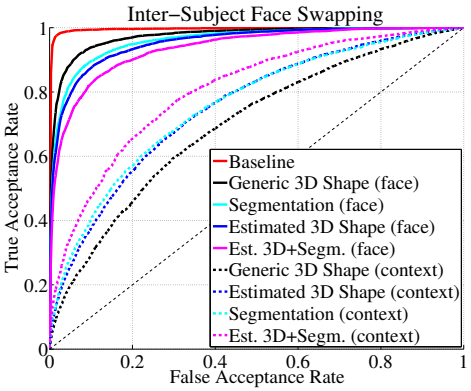

Figure 8: *Inter-subject swapping ROC curves.* Ablation study for the two experiments. Baseline shown in red.

perceived subject identity remains unchanged, supporting our *Assumption 2*.

In general, accuracy drops by ∼1%, with a similar nAUC compared to the use of original LFW images. This slight drop suggests that our swapping between different images of the same subject does not alter apparent facial identities.

V. Conclusions

We describe a simple face swapping method which is robust enough to allow for large scale, quantitative tests. From these tests, several key observations emerge. **(1)** State of the art face segmentation can be obtained with a standard segmentation network, provided that the network is trained on rich and diverse examples. **(2)** Collecting such examples is easy. **(3)** Both face and context play important roles in recognition. We offer quantitative support for the two decades old claim of Sinha and Poggio [35]. **(4)** Better swapping leads to more facial changes and a drop in recognition. Finally, **(5)**, 3D face shape estimation better blends the two faces together and so produces less recognizable faces.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA 2014-14071600011. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon. TH was also partly funded by the Israeli Ministry of Science, Technology and Space

References

[1] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec. Creating a photoreal digital actor: The digital emily project. In *Conf. Visual Media Production*, pages 176–187. IEEE, 2009.

[2] V. Axelrod and G. Yovel. External facial features modify the representation of internal facial features in the fusiform face area. *Neuroimage*, 52(2):720–725, 2010.

[3] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proc. Int. Conf. Comput. Vision*, December 2015.

[4] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: automatically replacing faces in photographs. *ACM Trans. on Graphics*, 27(3):39, 2008.

[5] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel. Exchanging faces in images. volume 23, pages 669–676. Wiley Online Library, 2004.

[6] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proc. Int. Conf. Comput. Vision*, pages 1513–1520. IEEE, 2013.

[7] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3D facial expression database for visual computing. *Trans. on Visualization and Computer Graphics*, 20(3):413–425, March 2014.

[8] T. Chen, P. Tan, L.-Q. Ma, M.-M. Cheng, A. Shamir, and S.-M. Hu. Poseshop: Human image database construction and personalized content synthesis. *Trans. on Visualization and Computer Graphics*, 19(5):824–837, 2013.

[9] G. Ghiasi, C. C. Fowlkes, and C. Irvine. Using segmentation to predict the absence of occluded parts. In *Proc. British Mach. Vision Conf.*, pages 22–1, 2015.

[10] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2010.

[11] T. Hassner. Viewing real-world faces in 3D. In *Proc. Int. Conf. Comput. Vision*, pages 3607–3614. IEEE, 2013. Available: www.openu.ac.il/home/hassner/projects/poses.

[12] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2015.

[13] M. Hrastnik. Faceswap code. Available: https://github.com/hrastnik/FaceSwap, 2017.

[14] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, UMass, Amherst, October 2007.

[15] X. Jia, H. Yang, K. Chan, and I. Patras. Structured semi-supervised forest for facial landmarks localization with face mask reasoning. In *Proc. British Mach. Vision Conf.*, 2014.

[16] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proc. Conf. Comput. Vision Pattern Recognition*. IEEE, 2014.

[17] I. Kemelmacher-Shlizerman. Transfiguring portraits. *ACM Trans. on Graphics*, 35(4):94, 2016.

[18] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark-A. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2015.

[19] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. *arXiv preprint arXiv:1611.09577*, 2016.

[20] M. Kowalski. Faceswap code. Available: https://github.com/MarekKowalski/FaceSwap, 2017.

[21] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[22] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua. Labeled faces in the wild: A survey. In *Advances in Face Detection and Facial Image Analysis*, pages 189–248. Springer, 2016.

[23] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009.

[24] G. Levi and T. Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Int. Conf. on Multimodal Interaction*. ACM, 2015.

[25] Y. Lin, Q. Lin, F. Tang, and S. Wang. Face replacement with large-pose differences. In *ACM Multimedia Conf.* ACM, 2012.

[26] S. Liu, J. Yang, C. Huang, and M.-H. Yang. Multi-objective convolutional learning for face labeling. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 3451–3459, 2015.

[27] C. Lu and X. Tang. Surpassing human-level face verification performance on lfw with gaussianface. *arXiv preprint arXiv:1404.3840*, 2014.

[28] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 2480–2487. IEEE, 2012.

[29] I. Masi, A. Tran, T. Hassner, J. T. Leksut, and G. Medioni. Do We Really Need to Collect Millions of Faces for Effective Face Recognition? In *European Conf. Comput. Vision*, 2016. Available www.openu.ac.il/home/hassner/projects/augmented_faces.

[30] S. Mosaddegh, L. Simon, and F. Jurie. Photorealistic face de-identification by aggregating donors face components. In *Asian Conf. Comput. Vision*, pages 159–174. Springer, 2014.

[31] M. A. Oikawa, Z. Dias, A. de Rezende Rocha, and S. Goldenstein. Manifold learning and spectral clustering for image phylogeny forests. *Trans. on Inform. Forensics and Security*, 11(1):5–18, 2016.

[32] P. Paysan, R. Knothe, B. Amberg, S. Romhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *Int. Conf. on Advanced Video and Signal based Surveillance*, 2009.

[33] P. Prez, M. Gangnet, and A. Blake. Poisson image editing. In *Proc. ACM SIGGRAPH Conf. Comput. Graphics*, 2003.

[34] S. Saito, T. Li, and H. Li. Real-time facial segmentation and performance capture from rgb input. In *European Conf. Comput. Vision*, pages 244–261. Springer, 2016.

[35] P. Sinha and T. Poggio. I think I know that face... *Nature*, 384(404), 1996.

[36] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, and J. Yang. Exemplar-based face parsing. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 3484–3491, 2013.

[37] A. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2017.

[38] L. Wolf, Z. Freund, and S. Avidan. An eye for an eye: A single camera gaze-replacement method. In *Proc. Conf. Comput. Vision Pattern Recognition*. IEEE, 2010.

[39] H. Yang, X. He, X. Jia, and I. Patras. Robust face alignment under occlusion via regional predictive power estimation. *Trans. Image Processing*, 24(8):2393–2403, Aug 2015.

[40] J. Zhu, L. Van Gool, and S. C. Hoi. Unsupervised face alignment by robust nonrigid mapping. In *Proc. Int. Conf. Comput. Vision*, pages 1265–1272. IEEE, 2009.

[41] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2016.