על סרגמטציה פיני, הח הפרים, ותפיסת פוני

עבורה תזה א ragazzo חכמל מחזור מוכשים לקבלי תואר
"מוסמך למידע" במדעי מחשב
באנוניברסיטה הפיטון
הotonin מחזוק למדעי מחשב

על יידי
יובל ירקני

העבורה הוכנה בהדרכאת של פורי"פ טל הנט"
탄크

אנחנומראים שהחלפת פנים היא תהליך דיפ_HOT(A)(במקום שלמרות מוערכת בודעת, אם כן, כמות שאריתם של פנים, או התערבות של פנים המתחברת במדורותTEL של הלחימ סנסצייתות מיתירות המודעות ביצורים מסוימים, כדי עד אומנות אל פנים, אם כן, אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות. אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשᤊים בכמויות פינות של ה lul. אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשSKTOPS כדי ל德拉 HOT של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשективים משﾎוות שלל ל德拉 HOT של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשSKTOPS כדי ל德拉 HOT של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשטלוו שלל ל德拉 HOT של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת מהירה ומדוייקת בפינות, אך גם שיעור לדינמיקה של פינות, ובשל ששתים ומשScrollPane שלל Labeled Faces in the wild (LFW) של אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת המוחה הפקות. אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת המוחה הפקות. אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת המוחה הפקות. אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת המוחה הפקות. אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת המוחה הפקות. אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת המוחה הפקות. אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת המוחה הפקות. אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת המוחה הפקות. אנחנומראים שבשתנו ירימה ביצורים מלאה יכולה '// להשיג מסגרת המוחה הפקות. אנחנומראים שבשתנו ירימה Бы Ше ШИШЭШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИШИश

ראיתי מומחה בזבב.
**List of Figures**

<table>
<thead>
<tr>
<th></th>
<th>Figure Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Inter-subject swapping.</td>
<td>4</td>
</tr>
<tr>
<td>3.1</td>
<td>Overview of our method.</td>
<td>8</td>
</tr>
<tr>
<td>3.2</td>
<td>Semi-supervised labeling and augmentation.</td>
<td>11</td>
</tr>
<tr>
<td>4.1</td>
<td>Qualitative segmentation results from the COFW (1-4) and LFW (5-8) data sets.</td>
<td>12</td>
</tr>
<tr>
<td>4.2</td>
<td>Qualitative LFW inter-subject face swapping results.</td>
<td>13</td>
</tr>
<tr>
<td>4.3</td>
<td>Comparison with [19] using the same pairs.</td>
<td>14</td>
</tr>
<tr>
<td>4.4</td>
<td>Face swapping failures.</td>
<td>14</td>
</tr>
<tr>
<td>4.5</td>
<td>Inter-subject swapping ROC curves.</td>
<td>17</td>
</tr>
<tr>
<td>4.6</td>
<td>Qualitative LFW intra-subject face swapping.</td>
<td>18</td>
</tr>
<tr>
<td>A.1</td>
<td>Qualitative intra-subject face swapping results.</td>
<td>24</td>
</tr>
<tr>
<td>B.1</td>
<td>Qualitative segmentation results from the COFW data set.</td>
<td>25</td>
</tr>
<tr>
<td>B.2</td>
<td>Qualitative segmentation results from the LFW data set.</td>
<td>26</td>
</tr>
<tr>
<td>C.1</td>
<td>Inter-subject face Swapping results.</td>
<td>27</td>
</tr>
<tr>
<td>C.2</td>
<td>Intra-subject face swapping results.</td>
<td>28</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>COFW [6] segmentation results.</td>
<td>13</td>
</tr>
<tr>
<td>4.2</td>
<td>Inter-subject face swapping.</td>
<td>16</td>
</tr>
<tr>
<td>4.3</td>
<td>Intra-subject face swapping.</td>
<td>17</td>
</tr>
</tbody>
</table>
Figure 1: Inter-subject swapping. LFW G.W. Bush photos swapped using our method onto very different subjects and images. Unlike previous work [4, 19], we do not select convenient targets for swapping. Is Bush hard to recognize? We offer quantitative evidence supporting Sinha and Poggio [40] showing that faces and context are both crucial for recognition.

Abstract

We show that even when face images are unconstrained and arbitrarily paired, face swapping between them is actually 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 to enable 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 well known 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 for machine vision systems.
1 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 [10, 35], it is actually far more important than this practice may suggest: Face swapping can also be used for preserving privacy [4, 34, 38], digital forensics [35] and as a potential face specific data augmentation method [33] especially in applications where training data is scarce (e.g., facial emotion recognition [26]).

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 [19]. 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 show how a rich image set with face segmentation labels can be generated with little effort by using motion cues and 3D data augmentation. The data we collect is used to train a standard FCN to segment faces, surpassing previous results in both accuracy and speed.

- **Face swapping pipeline.** We describe a pipeline for face swapping and show that our use of improved face segmentation and robust system components leads to high quality results even under challenging unconstrained conditions.

- **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 design two test protocols based on the LFW benchmark [15] to test how intra- and inter-subject face swapping affects face verification.

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 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 [40] in their well-known Clinton-Gore illusion, we are unaware of previous quantitative reports on how this applies to machine face recognition.

For code, deep models and more information, please see our project webpage.¹

¹www.openu.ac.il/home/hassner/projects/faceswap
2 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 [32] which segment individual facial regions (e.g., eyes, mouth) but not the entire face. An example based method was proposed in [41]. More recently, [11] 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 [29] a network was trained to simultaneously segment multiple facial regions, including the entire face. This method was used in the face swapping method of [19], but can be slow. The very recent method of [39] recently outperformed [11] on COFW as well as reported real-time processing speeds by using a deconvolutional neural network. We use a FCN for segmentation, proposing novel training data collection and augmentation methods to obtain challenging training examples.

Face swapping. Methods for swapping faces were proposed as far back as 2004 by [5] with fully automatic techniques described nearly a decade ago in [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, 9, 28, 34, 38]. Since then, however, many of their applications seem to come from recreation [19] or entertainment (e.g., [1, 44]).

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, 19]. 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, 27, 28], by fitting 3D Morphable Face Models (3DMM). Others instead estimate dense 2D active appearance models [9, 46]. This is presumably done in order to correctly map textures across different individual facial shapes.

Finally, deep learning was used to transfer faces [22], as if they were styles transferred between images. This method, however, requires the network to be trained for each source image and thus can be impractical in many applications.
3 Swapping faces in unconstrained images

Figure 3.1: Overview of our method. (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. 3.1). We show the 3DMM regressed by [42] but our tests demonstrate that a generic shape often works equally well. (c) Our face segmentation of Sec. 3.2 (red) overlaid on the projected 3D face (gray). (d) Source transferred onto target without blending, and the final results (e) after blending (Sec. 3.3).

Fig. 3.1 summarizes our face swapping method. When swapping a face from a source image, $I_S$, to a target image, $I_T$, we treat both images the same, apart from the final stage (Fig. 3.1(d)). Our method first localizes 2D facial landmarks in each image (Fig. 3.1(b)). We use an off-the-shelf detector for this purpose [18]. Using these landmarks, we compute 3D pose (viewpoint) and modify the 3D shape to account for expression. These steps are discussed in Sec. 3.1.

We then segment the faces from their backgrounds and occlusions (Fig. 3.1(c)) using a FCN trained to predict per-pixel face visibility (Sec. 3.2). We describe how we generate rich labeled data to effectively 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 image (Sec. 3.3).

3.1. Fitting 3D face shapes

To enrich our set of examples for training the segmentation network (Sec. 3.2) 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. 3.3). We experimented with two alternative methods of obtaining these 3D shapes.

The first, inspired by [14] 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 [42]. 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) [36] to represent faces and the 3DDFA Morphable Model [47] for expressions. These are
both publicly available 3DMM representations. More specifically, a 3-D face shape \( V \subset \mathbb{R}^3 \) is modeled
by combining the following independent generative models:

\[
V = \hat{v} + W_S \alpha + W_E \gamma. \tag{3.1}
\]

Here, vector \( \hat{v} \) is the mean face shape, computed over aligned facial 3D scans in the Basel Faces collec-
tion and represented by the concatenated 3D coordinates of their 3D points. When using a generic face
shape, we use this average face. Matrices \( W_S \) (shape) and \( W_E \) (expression) are principle components
obtained from the 3D face scans. Finally, \( \alpha \) is a subject-specific 99D parameter vector estimated sepa-
rately for each image and \( \gamma \) 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 \( \alpha \) using the deep network of [42]. They jointly
estimate 198D parameters for face shape and texture. Dropping the texture components, we obtain \( \alpha \)
and back-project the regressed face by \( \hat{v} + W_S \alpha \), 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,
\( p = \{p_i\} \subset \mathbb{R}^2 \), 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 \( \tilde{V} = \{\tilde{V}_i\} \) by:

\[
\tilde{V} \approx f(\hat{v}) + f(W_S) \alpha, \tag{3.2}
\]

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,
\( p_i \leftrightarrow \tilde{V}_i \), between detected facial features and their corresponding points on the 3D shape. Similarly
to [13], we use these correspondences to estimate 3D pose, computing 3D face rotation, \( R \in \mathbb{R}^3 \), and
translation vector \( t \in \mathbb{R}^3 \) using the EPnP solver [25].

Following pose estimation, we estimate the expression parameters in vector \( \gamma \) by formulating expres-
sion estimation as a bounded linear problem:

\[
\delta_R \left( P(R, t)(f(\hat{v}) + f(W_S) \alpha + f(W_E) \gamma) \right) = \delta_R(p),
\]

with \( |\gamma_j| \leq 3 \sigma_j \quad \forall \ j = \{1 \ldots 29\} \) \hspace{1cm} \tag{3.3}

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

3.2. 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., [39]). 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: [39], for example, used manually segmented LFW faces and the semi-automatic segmentations of [7]. These labels were costly to produce and limited in their variability and number. We obtain 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.

**FCN architecture.** We used the FCN-8s-VGG architecture, fine-tuned for segmentation on PASCAL by [30]. Following [30], we fuse information at different locations from layers with different strides. We refer to [30] for more details on this.

**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 [21]. 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 [12]. 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 [18] 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 [12] and adding or removing them from the face segmentation using simple mouse clicks. Fig. 3.2(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 from each 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. 3.1. 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 [18] 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 [39] by overlaying hand images at various positions on our example images. Hand images were taken from the egohands dataset of [3]. Fig 3.2(b) shows a synthetic hand augmentation and Fig 3.2(c) a sunglasses augmentation, along with their resulting segmentation labels.

### 3.3. Face swapping and blending

Face swapping from a source $I_S$ to target $I_T$ proceeds as follows. The 3D shape associated with the source, $V_S$, is projected down onto $I_S$ using its estimated pose, $P(R_S, t_S)$ (Sec. 3.1). We then sample
the source image using bilinear interpolation, to assign 3D vertices projected onto the segmented face (Sec. 3.2) with intensities sampled from the image at their projected coordinates.

The shapes for both source and target, $V_S$ and $V_T$ correspond in the indices of their vertices. We can therefore directly transfer these sampled intensities from all vertices $v_i \in V_S$ to $v_i \in V_T$. This provides texture for the vertices corresponding to visible regions in $I_S$ on the target 3D shape. We now render $V_T$ onto $I_T$, using the estimated target pose $(R_T, t_T)$, masking the rendered intensities using the target face segmentation (see Fig. 3.1(d)). Finally, the rendered, source face is blended-in with the target context using an off the shelf method [37].
4 Experiments

We performed comprehensive experiments in order to test our method, both qualitatively and quantitatively. Our face swapping method was implemented using MatConvNet [43] for segmentation, DLIB [20] for facial landmark detection and OpenCV [16] for all other image processing tasks. Run-times 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. We emphasize again that unlike previous work our implementation is publicly available.

4.1. Face segmentation results

Figure 4.1: Qualitative segmentation results from the COFW (1-4) and LFW (5-8) data sets.

Qualitative face segmentation results are provided in Fig. 3.1 and 4.1, visualized following [39] 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 [11], testing the 507 face photos in the COFW dataset [6]. Previous methods included the regional predictive power (RPP) estimation [45], Structured Forest [17], segmentation-aware part model (SAPM) [11], the deep method of [29], and [39]. Note that Structured Forest [17] and [39] used respectively 300 and 437 images for testing, without reporting which images were used. Result for [29] 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 [17], as well as two metrics from [17]: global – overall percent of correctly labeled pixels – and ave(face), the average face pixel recall. Tab. 4.1 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 [39] and measure run time for [29] ourselves. Other run times were reported in previous work.
<table>
<thead>
<tr>
<th>Method</th>
<th>mean IOU</th>
<th>Global ave(face)</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Struct. Forest [17]</td>
<td>–</td>
<td>83.9</td>
<td>88.6</td>
</tr>
<tr>
<td>RPP [45]</td>
<td>72.4</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SAPM [11]</td>
<td>83.5</td>
<td>88.6</td>
<td>87.1</td>
</tr>
<tr>
<td>Liu et al. [29]</td>
<td>72.9</td>
<td>79.8</td>
<td>89.9</td>
</tr>
<tr>
<td>Saito et al. [39] +GraphCut</td>
<td><strong>83.9</strong></td>
<td>88.7</td>
<td>92.7</td>
</tr>
<tr>
<td>Us</td>
<td>83.7</td>
<td><strong>88.8</strong></td>
<td><strong>94.1</strong></td>
</tr>
</tbody>
</table>


4.2. Qualitative face-swapping results

We provide face swapping examples produced on unconstrained LFW images [15] using randomly selected targets in Fig. 1, 3.1, and 4.2. 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. 4.3 shows a qualitative comparison with the very recent method [19] using the same source-target pairs. We note that [19] used the segmentation of [29] which we show in Sec. 4.1 to perform worst than our own. This is qualitatively evident in Fig. 4.3 by the face hairlines. Finally, Fig. 4.4 describes some typical failure cases and their causes.

Figure 4.2: Qualitative LFW inter-subject face swapping results. Examples were selected to represent extremely different poses (4,7,8), genders (1,2,7,8), expressions (1,7), ethnicities (1,3,6,8), ages (3-8) and occlusions (1,5).
4.3. Quantitative tests

Similarly to previous work, we offer qualitative results to visualize the realism and quality of our swapped faces (Sec. 4.2). Unlike previous work, 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 the context of the face (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 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 [40] 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 unnecessary changes to facial appearances.

To test these assumptions, we produce modified (face swapped) versions of the LFW benchmark [15]. 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 [31]. 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 [24].

**Face verification.** We use the ResFace101 [33] 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 [21]. 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 (e.g., Fig. 4.4), 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 \((I_1^i, I_2^i)\) be the \(i\)-th LFW test image pair. We produce \(\hat{I}_1^i\), the swapped version of \(I_1^i\), by randomly picking another LFW subject and image from that subject as a target, taking \(I_1^i\) as the source. We then do the same for \(I_2^i\) to obtain \(\hat{I}_2^i\).

Ostensibly, we can now measure recognition accuracy using the new pairs \((\hat{I}_1^i, \hat{I}_2^i)\). 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 \((\hat{I}_1^i, I_2^i)\), the other on pairs \((I_1^i, \hat{I}_2^i)\). We then report the average results for both trials. We refer to these tests as *face preserving* tests, as swapping preserves the face, transferring it onto a new context.

We also performed *context preserving* tests. These tests use benchmark image pairs as targets and not sources as before. Thus the faces in these images preserve the context of the original LFW images, rather than 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 [23]. 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 their tests which used black boxes to occlude faces, our images contain faces of other subjects swapped in place of the original faces, and so our images are more challenging than theirs.

**Inter-subject swapping results.** We provide verification results for both face preserving and context preserving inter-subject face swapping in Tab. 4.2 and ROC curves for the various tests in Fig. 4.5. Our results include ablation studies, showing accuracy with a generic face and no segmentation (*Generic*), with an estimated 3D face shape (Sec. 3.1) 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. 4.2 (bottom) are consistent 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 $\sim 8\%$ better accuracy than using our entire pipeline. Importantly, just by estimating 3D face shapes, accuracy drops by $\sim 3.5\%$ compared to using a simple generic face shape.

Unsurprisingly, the context preserving results in Tab. 4.2 (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. These results, obtained using different components of our system, are provided in Tab. 4.3 and Fig. 4.5. Example intra-subject face swap results are provided in Fig. 4.6. These show that even under extremely different viewing conditions, perceived subject identity remains unchanged, supporting our Assumption 2.

In general accuracy drops by 1%, with a similar nAUC, compared to the use of the original, unchanged LFW images. This very slight drop suggests that our swapping does not obscure apparent face identities, when performed between different images of the same subject. Moreover, in such cases, accuracy for estimated vs. generic 3D shapes is nearly identical.
Method & 100%-EER & Acc. & nAUC \\
--- & --- & --- & --- \\
Baseline & 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 \\

Table 4.3: Intra-subject face swapping. Ablation study.

Figure 4.5: Inter-subject swapping ROC curves. Ablation study for the two experiments. Baseline shown in red.
Figure 4.6: *Qualitative LFW intra-subject face swapping.* Faces remain recognizable despite substantially different view settings.
5 Conclusions

We describe a novel, simple face swapping method which is robust enough to allow for large scale, quantitative testing. From these tests, several important observations emerge. (1) Face segmentation state of the art speed and accuracy, outperforming methods tailored for this task, 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 faces and their contexts play important roles in recognition. We offer quantitative support for the two decades old claim of Sinha and Poggio [40]. (4) Better swapping, (e.g., to better mask facial spoofing attacks on biometric systems) 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 source faces.
Bibliography


[40] P. Sinha and T. Poggio. I think I know that face... *Nature*, 384(404), 1996. 4, 6, 14, 19


Additional results
A Additional intra-subject qualitative results

Figure A.1: Qualitative intra-subject face swapping results.
B Additional segmentation results

Figure B.1: Qualitative segmentation results from the COFW data set.
Figure B.2: Qualitative segmentation results from the LFW data set.
C  Qualitative ablation results

![Inter-subject face Swapping results. Qualitative ablation study.](image)

Figure C.1: *Inter-subject face Swapping results.* Qualitative ablation study.
Figure C.2: Intra-subject face swapping results. Qualitative ablation study.