Bringing Portraits to Life

HADAR AVERBUCH-ELOR, Tel-Aviv University
DANIEL COHEN-OR, Tel-Aviv University
JOHANNES KOPF, Facebook
MICHAEL F. COHEN, Facebook

We present a technique to automatically animate a still portrait, making it possible for the subject in the photo to come to life and express various emotions. We use a driving video (of a different subject) and develop means to transfer the expressiveness of the subject in the driving video to the target portrait. In contrast to previous work that requires an input video of the target face to reenact a facial performance, our technique uses only a *single* target image. We animate the target image through 2D warps that imitate the facial transformations in the driving video. As warps alone do not carry the full expressiveness of the face, we add fine-scale dynamic details which are commonly associated with facial expressions such as creases and wrinkles. Furthermore, we hallucinate regions that are hidden in the input target face, most notably in the inner mouth. Our technique gives rise to *reactive profiles*, where people in still images can automatically interact with their viewers. We demonstrate our technique operating on numerous still portraits from the internet.

CCS Concepts: • **Computing methodologies** → *Animation*;

Additional Key Words and Phrases: face animation, facial reenactment

ACM Reference format:
Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. 2017. Bringing Portraits to Life. *ACM Trans. Graph.* 36, 4, Article 196 (November 2017), 13 pages.
https://doi.org/10.1145/3130800.3130818

Fig. 1. Given a single image (top row), our method automatically generates photo-realistic videos that express various emotions. We use driving videos of a different subject and mimic the expressiveness of the subject in the driving video. Representative frames from the videos are displayed above. ©Unsplash photographers Lauren Ferstl, Brooke Cagle, Guillaume Bolduc, Ilya Yakover, Drew Graham and Ryan Holloway.

1 INTRODUCTION

Few objects convey as large a range of depth and meaning as the human face. Facial expressions in humans convey not just major emotions, but through subtle variations, a rather nuanced view into the emotional state of a person, for example, a sad smile, blushing, etc. (e.g., much has been said about the smile of Mona Lisa).

In this work, we are interested in animating faces in human portraits, and in particular controlling their expressions. To avoid crossing into the "uncanny valley", previous facial animation techniques usually assume the availability of a video of the target face, which exhibits variation in both pose and expression [Dale et al. 2011; Garrido et al. 2014; Thies et al. 2016]. An input video, or even an image collection of the target face (e.g., [Cao et al. 2016]), allows for an accurate 3D face reconstruction, over which face textures are mapped and manipulated.

In contrast to previous work we use as input only a *single* image of a target face to animate it. This makes our method more widely applicable, to the near endless supply of portrait or selfie images on the internet. We animate the single target face image from a driving video, allowing the target image to *come alive* and mimic the expressiveness of the subject in the driving video. While most previous work restrict themselves to only the face region, within limits, our method animates the full head and upper body.

We animate the target image by a series of warps that imitate the facial transformations in the driving video. Like in previous works (e.g., [Fried et al. 2016; Leyvand et al. 2008; Yang et al. 2011]), we manipulate the face by lightweight 2D warps. We are able to create moderate head movement while maintaining the high realism of the input 2D image without converting and projecting the image to 3D.

In our work, we establish a correspondence between the target image and driving video frames by utilizing the effectiveness of facial landmarks detection and tracking techniques, and expanding these facial correspondences to span the entire image and over time. As warps alone do not carry the full expression of the face, we add fine-scale details such as wrinkles and creases that are commonly associated with facial expressions, and hallucinate regions that are hidden in the input target face, most notably in the inner mouth. Figure 1 shows results of expressions which were transferred to a single target image on the left. (The reader is encouraged to see the videos included in the supplementary material).

We thank Peter Hedman, Noa Fish, Tal Hassner and Amit Bermano for their insightful comments and suggestions. We also thank Ohad Fried, Justus Thies, Matthias Niessner, Pablo Garrido, and Christian Theobalt for providing us with comparisons to their techniques. This work is partially supported by the Israeli Science Foundation, research program (1790/12 and 2366/16).

As our results illustrate, our technique enables bringing a still portrait to life, making it seem as though the person is breathing, smiling, frowning, or for that matter any other animation that one wants to drive with. We apply our technique on highly varying facial images, including internet selfies, old portraits and facial avatars. Additionally, we demonstrate our results in the context of *reactive profiles* – a novel application which resembles the moving portraits from Harry Potter's magical world, where people in photographs move, wave, etc. By discovering portions of driving videos which contain a number of different emotions, our technique provides the means for the target image to *react* to stimulii. Our main contributions are:

- The ability to bring life to a single still portrait through 2D warps and generate a video sequence which maintains the realism of the input image.
- A method to continuously transfer fine-scale details including wrinkles and creases, while avoiding *outlier* wrinkles that are caused by cast shadows or misalignment between the warped video frames.
- A method to seamlessly transfer hidden regions, e.g., the mouth interior, when necessary.
- A new *reactive profile* application, enabled from a single input image.

2 RELATED WORK

Our method takes a single target image of a neutral face in frontal pose and generates a video that expresses various emotions. Previous works [Blanz and Vetter 1999; Breuer et al. 2008] addressed face manipulation from a single image, but do not focus on generating an animated video. Other works [Thies et al. 2016; Vlasic et al. 2005] manipulate or reenact a facial performance, but these assume the availability of a video of the target face. In what follows, we elaborate on the most closely related works.

Most prior works require more than a single target image of a face to automatically manipulate it. A video-to-image facial retargeting application was previously introduced in Cao et al. [2014], but their method is not automatic and requires some user interaction. Facial editing using deep networks was introduced in Yeh et al. [2016]. The method of Liu et al. [2001] enables transferring the fine-scale details of one person's changed expression to a neutral target image. In our work, we extend their technique to accommodate more general image pairs and a stable video output. Recently, Fried et al. [2016] presented a technique to manipulate the camera viewpoint from a single input image. Their method enables modifying the apparent relative pose and the distance between the camera and the subject. Other works, such as [Hassner et al. 2015], specifically address the problem of face frontalization, as it is extremely beneficial for facial recognition [Ding and Tao 2016].

In their seminal work, Blanz and Vetter [1999] fit a 3D morphable model to a single input image, texture-map a face image onto a 3D mesh, and parametrically change its pose and appearance. In a follow up work, Blanz et al. [2003] enabled animating a single image, but focus on the mouth region. Later works, e.g. [Breuer et al. 2008], extended the 3D morphable model technique to allow for an automatic reconstruction pipeline. However, as noted by Piotraschke and Blanz [2016], when only a single image is provided, plausible reconstructions often require a manual initialization. Furthermore, the realism of the manipulated faces using these techniques in general is often lacking, as they cannot extract fine details since they are not spanned by the principal components. To maintain the realism of the input image, one should avoid the projection of the image onto a 3D model. This claim holds for faces as well as for the whole body (e.g., [Zhou et al. 2010]). The commercial system FaceApp [1] can generate a smiling still image from a neutral input image. In the supplementary material, we demonstrate some output stills generated by their system.

Facial manipulation techniques that require an input video of the target face are by far more common. Vlasic et al. [2005] edit the 3D mesh of the target face according to the expression parameters. Li et al. [2012] utilize a facial performance database of the target face. Dale et al. [2011] use a 3D morphable model for face reeanactment using a source and target video. Garrido et al. [2014] present an automatic method for face replacement in video, but unlike Dale et al., they only replace the actor's facial region, and keep the hair and the rest of the head and the upper body of the source video. The problem of face swapping was also introduced in the context of an input target image, for example in [Korshunova et al. 2016]. Recently, Thies et al. [2016] presented a real-time facial reenactment of a target video sequence. Unlike our animation of a single static image, their work assumes that the target video contains rich and sufficient data to synthesize a plausible reenactment. In our work, we do not reenact the facial expression only, but also the respective head motion. The technique of Kemelmacher et al. [2010] is also related to ours, but more in the context of image retrieval. Other video techniques, such as [Garrido et al. 2015], focus specifically on transferring lip motion to an existing target video.

There are other works that address facial expression editing in video, but do not explicitly follow another performer. Some examples include [Kuster et al. 2012] who aim at retargeting the gaze of a streaming video and [Ganin et al. 2016] who manipulate the gaze of a single image. Some works focus, for example, on synthesizing a realistic inside of the mouth (e.g., [Kawai et al. 2013, 2014]). Bai et al. [2013] edit a facial video performance in an expression-preserving manner, removing undesired large-scale motion. Yang et al. [2012] magnify (or suppress) the provided expression in a target video. In contrast, we create new photo-realistic expressions which are significantly different from the input image. Recently, Masi et al. [2016] addressed the problem of data augmentation by adding expression variation in the form of a single local expression control (closing the mouth). This expression manipulation helps in augmenting training data, without necessarily generating a plausible realistic expression.

Some previous works address transferring expressions from a user to a facial avatar (e.g., [Saragih et al. 2011]) or from one facial avatar to another (e.g., [Chuang and Bregler 2005]). Although our method enables animating non photo-realistic faces, the main challenge in our work is to maintain a high degree of realism of a human face. Unlike animating and manipulating the full body (e.g., [Hornung et al. 2007; Zhou et al. 2010]), in facial animation, we, as humans,

[1] https://www.faceapp.com

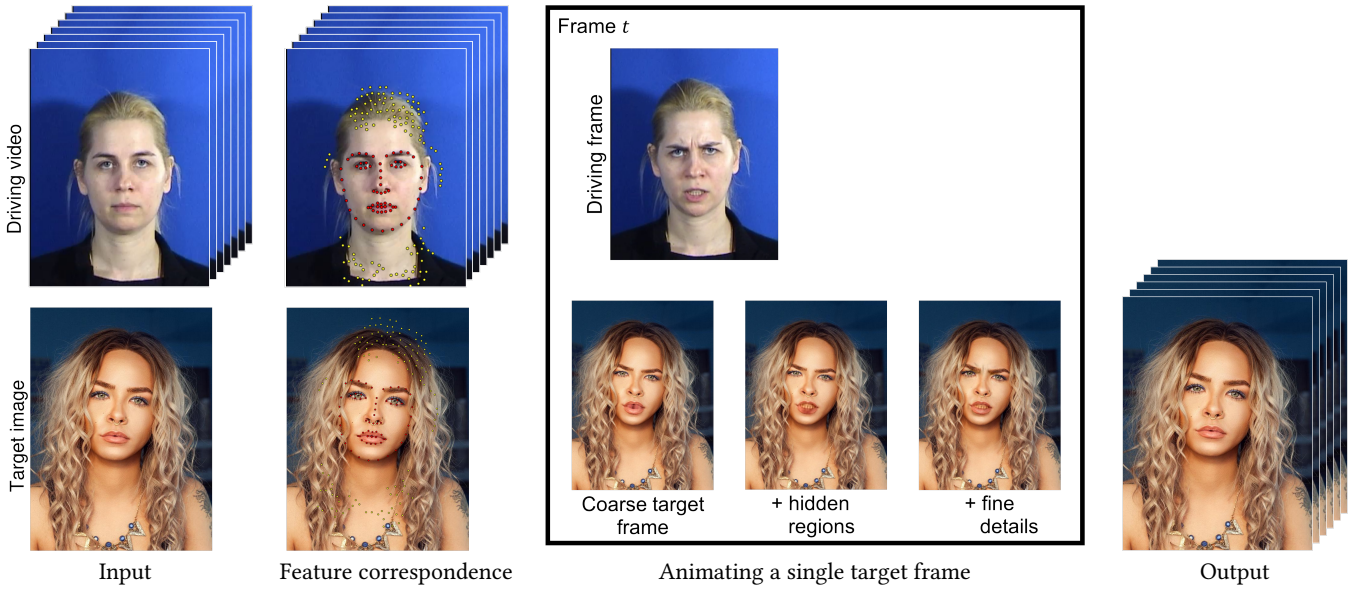


Fig. 2. An overview of our method. Given an input target image and a driving video, we extract and track facial and non-facial features in the driving video (colored in red and yellow, respectively) and compute correspondences to the target image. To generate the animated target frames, we perform a 2D warping to generate a coarse target frame, followed by a transfer of hidden regions (i.e., the mouth interior) and fine-scale details. ©Unsplash photographer Atikh Bana.

are extremely sensitive to the finest nuances. To do so, we merely utilize 2D tracked features in the driving videos, bypassing the need for a precise tracking procedure such as the ones presented in Cao et al. [2015] or Saito et al. [2016].

3 OVERVIEW

Our method uses as input a single image of a target face in neutral-frontal pose, as well as a video of a different face that drives the animation of the target image. An overview of our method is illustrated in Figure 2. We animate the target image through a series of 2D warps that imitate the facial transformations in the driving video. These warps are controlled by a set of sparse correspondences between the target face and the face in the driving video frames. Many methods have been proposed for detecting *fiducial points* on faces at fixed standard locations (eyes, mouth corners, etc., see red points in Figure 2). However, to bring the full head of the target image to life we need to allow the entire head to move and change its pose. Thus, we track additional points outside the face region to help guide the overall head.

Moreover, geometric warps alone do not encode the full range of changes a face undergoes when the expression changes. In addition there are many fine-scale changes, such as self-shadowing in wrinkles and creases that are necessary to convey the full expression of the face in the driving video. Furthermore, by assuming a neutral target face, we implicitly assume a closed mouth of the target face. Therefore, the inner mouth of the target face is hidden, and thus we need to hallucinate the appearance in this region if the mouth opens in the driving video.

To meet these requirements, we develop the following technical solutions:

Correspondence expansion. We utilize the high-fidelity of facial landmarks detection and tracking, and expand the correspondences in the facial region to correspondences that span the entire image and over time. These additional landmarks are illustrated in yellow in Figure 2. The augmented set of corresponding points allow tracking and changing the pose of the target head to follow and imitate the one in the driving video (see Section 4).

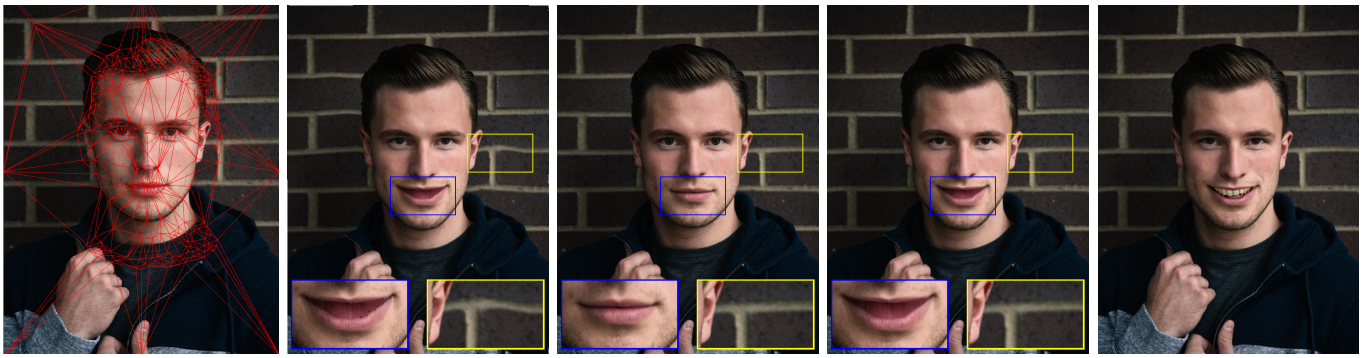
Confidence-aware warping. We extrapolate the sparse set of correspondences to a dense vector field over the entire image. See the illustration of the warped target at frame t. We distinguish between the highly-confident facial region and the rest of the image, where we have no guarantees on the quality or even the quantity of the corresponding points, and smooth the vector field accordingly (see Section 4).

Hidden region transfer. When needed (i.e., the mouth is open in the driving video), we transfer the mouth interior to the animated target frame. The composite retains as many details as possible from the target image as only the mouth interior, and not the lips themselves, are transferred to the animated frame. Note the hallucinated teeth of the animated target face at frame t (see Section 5).

Detection of inlier wrinkles. To generate a realistic expression, we transfer creases and wrinkles in the facial region. We detect and avoid illumination changes that are caused by cast shadows or misalignments between the warped video frames (see Section 6).

4 COARSE TARGET VIDEO SYNTHESIS

Our input consists of a target image, t^*, and a driving video S, which contains a series of frames s_i. Note that our notation assigns lowercase letters to images and frames and uppercase letters to



(a) Input target image (b) No smoothing (c) Constant smoothing (d) Our warped frame (e) Our final frame

Fig. 3. Confidence-aware warping. (a) The target image to be warped, with the triangulation mesh on top of it. (b) Without smoothing, discontinuities can be observed in the non-facial region, e.g., in the region inside the yellow rectangle. (c) A constant blur kernel diminishes the animated facial expression. As can be seen inside the blue rectangle, the mouth is no longer smiling. (d) Our confidence-aware blurring kernel keeps the facial details and smooths out the discontinuities. (e) The animated frame on the right is our final result (after transferring hidden regions and fine-scale details). ©Unsplash photographer Jimmy Bay.

videos. Our goal is to synthesize a new video T, constructed from frames t_i, that follows the motion of S, but maintains the identity from t^*. We assume that t^* is an image of a neutral face and that there is a frame, s^*, in S, with a neutral expression.

Each frame in the synthesized sequence $t_i \in T$ corresponds to a frame in the driving video $s_i \in S$. To ease notations, we assume the neutral frame of the driving video, s^*, to be the first frame (s_0), but this frame can be selected as well. Each frame $t_i \in T$ is generated by warping t^* according to a sparse set of control points $\mathbf{p}_i^t$, which mimic the offsets of the corresponding set of control points $\mathbf{p}_i^S$ defined in the driving video.

Since the faces in t^* and s_0 are not aligned, we first compute an aligning transformation, ϕ, between t^* and s^* that compensates for the misalignment. For each frame, $t_i \in T$, t^* is warped by any changes in the positions of the points in the driving video, $(\mathbf{p}_i^S - \mathbf{p}_0^S)$ after the transformation, ϕ, that aligns the neutral frames. Putting it all together:

$$\mathbf{p}_i^t = \mathbf{p}_0^t + \phi \cdot \left(\mathbf{p}_i^S - \mathbf{p}_0^S\right) \tag{1}$$

The alignment and subsequent offsets are computed over a set of control points. The control points that we use consist of two sets, illustrated in Figure 2 in red and yellow points. The red points are the fiducial points on the faces. Various methods have been proposed for detecting fiducial points on faces. We use the implementation available in [King 2009] which automatically detects and tracks 68 facial landmarks. This provides us with 68 corresponding landmarks which are located in the chin, mouth, nose, eye and eyebrow regions.

The aligning transformation ϕ is defined once by estimating a similarity transformation between the target image t^* and the neutral frame s^*. We estimate this transformation using least-squares by approximating a rotation and scale between the landmarks located in the eye regions and the tip of the nose.

To modify the entire image, and not just the facial region, some correspondences outside the facial region are required. However, there are no consistent features away from the facial region in both the target and driving video. Thus, a robust correspondence technique cannot be reliably obtained. We therefore instead track points in the driving video and *hallucinate* the corresponding locations of points in the target image. To track points in the driving video, we use a simple optical flow tracker [Bouguet 2001]. We then hallucinate the corresponding pixel locations in the target image using the aligning transformation ϕ. This peripheral set of control points are illustrated with yellow points in the Figure 2. The peripheral set also includes points along the image boundary that do not move throughout the animation to help fix the background in place. The fiducial points together with the peripheral points form the augmented set of control points used to warp t^* at each frame.

To interpolate the offsets of the sparse set of control points to a dense warp field for the entire image, we use a Delauney triangulation of the control points (see Figure 3(a)), which is computed on s^*. Corresponding triangles on t^* define a simple piece-wise linear interpolation, which works well inside the facial region, where the points are relatively dense, but may cause noticeable discontinuities outside the facial region.

To alleviate this issue, we smooth the dense warp field with respect to a confidence associated with each point. Inside the facial region, where we have reliable correspondences, smoothing is unnecessary and may take away from the desired facial expression (e.g., the warped mouth in Figure 3(c) is no longer smiling). Outside the facial region, we smooth the warp field by convolving it with a disk whose radius increases away from the facial region. For efficiency, we use 10 different blurring kernels with radii in the range $\left[0, 0.05 \cdot S_{diag}\right]$, where S_{diag} is the size of the image diagonal. We then use inverse mapping to calculate the origin of each pixel in the new animated target frame. As can be observed in Figure 3(d) the spatially-variant warp keeps the facial details, and at the same time smooths out the discontinuities.

5 TRANSFERRING HIDDEN REGIONS

When the driving video's subject opens its mouth, the interior is revealed which does not exist in the target image. We thus transfer this region from each driving frame to the target video frame.

Target frame t_i | Warped video frames s^* and s_i | R_i and result by Liu et al. [2001] | Our refined R_i and result

Fig. 4. Our fine-scale details transfer procedure. For each target frame t_i, we use the aligned video frame s^* and s_i to obtain the analogous expressive form of t_i. As the figure demonstrates, the full ERI contains various artifacts. Our refined version yields a considerably cleaner result. ©Unsplash photographer Seth Doyle.

To achieve a natural composite, we first align the driving frame $s_i \in S$ with the warped target frame $t_i \in T$ using the warping procedure described in the previous section. We then generate a compatible target mouth using Poisson blending [Pérez et al. 2003] to match color at the transition.

The crop region is determined according to the facial landmarks on the mouth exterior. It is important to further ensure that these landmarks fall within the lip region (and not on the skin). Thus, we perform a morphological erosion on this mask. The disk radius is $0.1 \cdot h_{mouth}$, where h_{mouth} is the height (in pixels) of the mouth in the target image t^*. We then alpha blend the compatible mouth into the target image. The region that is transferred is determined according to the facial landmarks on the mouth interior, and therefore, the lips themselves are not transferred to the target face.

We only transfer the mouth interior when its size is significantly bigger than in the input target image (otherwise we assume that it is not hidden). Denote the area (in pixels) of the mouth interior in the target image by a^*_{mouth}, then the mouth interior is transferred when $a^t_{mouth} > 2 \cdot a^*_{mouth}$. To achieve a temporally smooth sequence, we linearly blend between the two mouth interiors when it is in the range $\left[a^*_{mouth}, 2 \cdot a^*_{mouth}\right]$.

6 TRANSFERRING FINE-SCALE DETAILS

In Section 4 we described how to warp the target image t^* to generate the initial series of animated frames $t_i \in T$. To fully generate a realistic animated expression, we need to augment the target frames with fine-scale shading changes extracted from the source video. These details include the shading induced by wrinkles around the eyes when we smile, or the creases alongside the smiling mouth.

Liu et al. [2001] presented a technique to compute an *expression ratio image* (ERI), which captures such illumination changes of one person's expression. An ERI computation requires three *aligned* images: two *unexpressive* images, denoted by I_a and I_b, one in the source and one in the target, and an *expressive* image of the source, denoted by $\tilde{I}_a$. The ratio image R, thus, approximates the illumination changes between the source pair:

$$R = \frac{f(\tilde{I}_a)}{f(I_a)}, \tag{2}$$

where the function $f(x)$ is some general function applied on the input image, for example, taking the luminance channel only. The unexpressive image in the target is then modified to obtain $\tilde{I}_b$ which is the analogous expressive form:

$$\tilde{I}_b = R \cdot I_b. \tag{3}$$

Therefore, for each frame t_i, we align the unexpressive driving frame $s^* \in S$ and the expressive driving frame $s_i \in S$ with the warped target frame $t_i \in T$. To do so, we warp s^* and s_i as described in the previous section (see the warped video frames in Figure 4). We can then compute the ratio image R_i and apply it to t_i.

As can be observed in Figure 4, applying the full ratio image R_i everywhere yields various undesired artifacts: (i) certain areas in the image may become saturated, (ii) it can include *outliers*, which are illumination changes that may be caused by shadowing from the nose or misalignments (see the shadow beside the nose in the figure), and (iii) it lacks temporal stability. Furthermore, undesired artifacts may appear in the hair and the background in general. Although our landmark detector can provide a partial facial mask, which includes the chin up to the eyebrows, the forehead region must be included as well to enrich the expressiveness of the face. In what follows, we elaborate on our solutions to these issues that provide the results seen in Figure 4 on the right-hand side.

Robustness to saturation. We have found the darkening inside the wrinkles to be the most essential aspect of the ratio images. Also, pixels in the ratio image that brighten, i.e., have value greater than 1, are often due to small misalignments of very dark regions near the nostrils for example.

To avoid saturation in the output image, we tune down the brightening effect by multiplying any pixel whose corresponding pixel in R_i is larger than 1 by a constant fraction. We use a multiplying factor of 0.01 in our implementation.

Facial region estimation. As the ERI is computed everywhere, we need to estimate the facial region to avoid changes outside it. Since we have 2D landmarks along the chin and the eyebrows, the only region that still remains to be estimated is the forehead. To estimate the location of the forehead, we first fit an ellipse to the points along the chin. This provides us with an initial estimate. The initial estimate is refined using a Grab-cut optimization [Rother et al. 2004], and the ratio multiplier is only applied inside the face region.

Outlier detection and elimination. Certain values in the ratio image are not due to momentary changes such as those that occur in

transient creases or wrinkles that are important in conveying emotion. Rather, they are due to cast shadows that may move slightly due to face motion. We would like to remove these effects in the ratio images. Some facial regions are more prone to cast shadows for which we do not want to apply the ratio image. The most notable example is probably the side of the nose, which is commonly shadowed by the nose itself. Another is the inside of the nostrils. We would like to eliminate the effects of such ratio values by setting them to 1, while allowing the transient ones to be used. We achieve this with an outlier detection method.

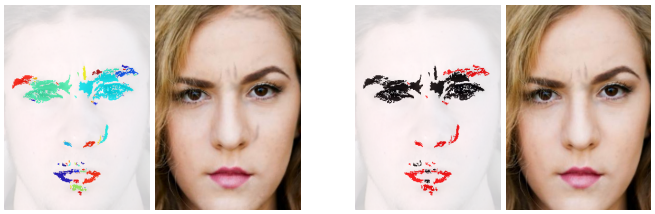
In this context, outliers are pixels with significant ratio values in a ratio image. In our implementation, values greater than 1.1 and less than $\frac{1}{1.1}$ are considered significant. To be an outlier, these values must also appear at approximately the same place in the neutral frame, s^*. To detect these, we developed a method which is conceptually similar to template matching. For each pixel in s_i, we examine a small $3x3$ neighborhood around it and perform a local search for the best matching neighborhood in s^* (within a $20x20$ region and allowing up to 30 deg rotation).

Rather than thresholding the best match per pixel, per frame, to determine outlier pixels, we do two things. First, we find connected components of significant ratio values. Second, we perform the outlier detection only once in a reference frame s_{ref}, the one that is the most distinct from the neutral frame, s^*. All pixels in the connected components in the reference frame that pass the threshold are then propagated to other frames where the corresponding pixels in the ratio images are set to 1. We will discuss each of these steps in the following paragraphs.

In our implementation, connected components are formed by 8-neighbor connected pixels with significant ratio values in R_i (see Figure 5, left-hand side). For each pixel in the connected component, we record the minimum within the $20x20$ search window, of the maximum RGB pixel difference in the $3x3$ neighborhood template. We then average these minima, and if the average is less than 5, we label the entire connected component as an outlier. Pixels in outlier components in R_i are all set to 1 thus removing their effect in the final image. See Figure 5 for an illustration of this procedure (note that the corresponding warped video frame is illustrated in Figure 4). As Figure 5 illustrates, most of the outlier regions are detected. However, some regions still remain. For example, the left brow and the left eye form a single connected component, obtaining an average which is above the threshold due to the changes that occur in the eye region.

To detect the expression reference frame, we estimate a similarity transformation $\phi_{S^* \rightarrow s_i}$ between the landmarks in the first frame and the landmarks of the current frame. We then compute the average L2 distance between the transformed facial landmarks. The frame with the largest distance is then picked as the most *not-neutral* frame and is used as the reference frame, s_{ref}, to detect the connected components of outliers.

We then propagate the footprint of the connected components to the rest of the video sequence. For each frame, we estimate a similarity transformation $\phi_{s_i \rightarrow s_{ref}}$ between the landmarks in the current frame and the landmarks of the reference frame. For each pixel that is close enough to an outlier (less than 20 pixels in our implementation), its R_i value is set to one.



Considering all regions | Eliminating outlier regions

Fig. 5. Outlier detection. Left: Considering all the regions with significant ratio values (each connected component is colored in a unique color) yields the expressive image on the left. Right: By eliminating the outlier regions (highlighted in red), we obtain a refined result, which does not contain, for example, the cast shadow beside the nose. ©Unsplash photographer Seth Doyle.

	Real Videos					**Animated Videos**				
	1	2	3	4	5	1	2	3	4	5
Anger	0.00	0.11	0.11	0.42	0.35	0.10	0.28	0.21	0.28	0.13
Fear	0.00	0.02	0.08	0.6	0.31	0.08	0.26	0.16	0.34	0.15
Happy	0.00	0.02	0.11	0.40	0.47	0.02	0.17	0.23	0.33	0.24
Surprise	0.00	0.03	0.13	0.26	0.57	0.12	0.33	0.19	0.25	0.11
Average	0.00	0.04	0.11	0.42	0.43	0.08	0.26	0.20	0.30	0.16

Table 1. The user study results. The rankings (1-5) signify low (very likely fake) to high (very likely real) scores.

Finally, to increase temporal stability for the sequence of ratio images, we convolve a 21 frametime wide temporal Gaussian filter over the aligned ratio images.

7 RESULTS AND EVALUATION

7.1 User study

We conducted a user study to quantitatively evaluate the quality of our results. The participants were presented with real and animated videos. We used videos from the MMI Facial Expression Database [Pantic et al. 2005; Valstar and Pantic 2010]. The database contains videos of persons expressing various emotions. We used videos of subjects that had a complete set of four emotions (anger, fear, surprise, happiness).

The videos contain a full temporal pattern, starting from a neutral face, through a series of frames leading to the respective emotion, and back to the neutral face. The animated videos were generated by selecting the first video frames to be the target images and driving these target images by one of the 3×4 (3 other subjects and 4 available emotions) driving videos.

The participants were presented with 24 randomly selected videos, eight of which are real. They were asked to rate them based on *how real* the animation looks. The 5-point Likert range of scores they were provided with are: *very likely fake, likely fake, could equally be real or fake, likely real, very likely real.* The participants were allowed only one viewing of each video, to evaluate their first impression, without having the option to repeat and look for problems. Thirty users in the age range 20-50 participated in the study (with a 1.5 male to female ratio). We did not normalize the differences in individual aptitude, as the differences were small.

The results of the user study are illustrated in Table 1. Overall, only 85% of the real videos were identified as such (were rated as either likely real or very likely real). It seems that the participants often over-analyzed the videos and convinced themselves that the real videos looked fake. Our animated videos were identified as real 46% of the time. The "happy" animations were perceived as the *most real* (identified as real 58% of the time), while the "surprise" animations were perceived as the *least real* (identified as real 37% of the time). We believe that our success in hallucinating real videos with our technique is surprisingly good, given the fact that humans in general are particularly sensitive to facial nuances, especially given that users were primed to look for fakery.

7.2 Pipeline evaluation

We demonstrate our technique on challenging casually-captured internet portraits (Figures 1-3,12,13).

To drive these casual portraits, we used videos from the MMI Facial Expression Database as well as videos that we captured on our own. In Figure 13, we illustrate a sample of driving video frames of subjects that we captured and the corresponding target animations. In the paper, we display only a representative image, but more results are included in the accompanying video and in the supplementary material.

To demonstrate the importance of the various stages in our pipeline, we provide intermediate results of our technique in Figure 12. In the supplementary material, we provide the intermediate results of the full animations. In what follows, we compare to alternative methods.

7.2.1 Warping-only comparison to previous work. To evaluate our confidence-aware warping technique, we compare against Fried et al. [2016]. In their dense geometry proxy they *manually* annotate three fiducial landmarks which is an overly meticulous task for an entire video. Thus, we compare against their 2D warping strategy by providing their technique with the sparse set of correspondences we compute according to Section 4. Figure 6 demonstrates the results on a few image pairs. While the differences are subtle, the advantage of our confidence-aware warping technique is visible especially along strong edges (see the highlighted regions). As the figure demonstrates, the results obtained using our warping technique are smoother. Furthermore, the final results of our method also include the fine-scale details and the missing regions, and not only the geometric warp which is demonstrated in the figure.

7.2.2 Reenactment evaluation. To evaluate our reenactment results, we compare our results to state-of-the-art video-to-video reenactment techniques. To meet their input requirements for a video, we replicate the single target image we use to form a static video. In the supplementary material, we include side-by-side comparisons of full sequences. Please refer to the full comparisons to better appreciate the differences.

Thies et al. [2016]: In their work, they construct a mouth dataset from the target input video. Since we start with a single image, all the target frames contain the same replicated closed-mouth, thus the mouth motion graph construction fails as it only contains a single cluster. Therefore, for the sake of a comparison, the authors

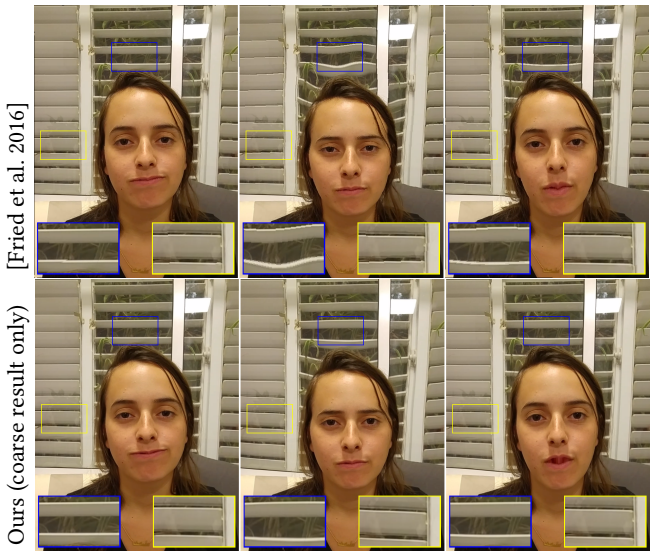


Fig. 6. Comparison against the image-based warping technique of Fried et al.[2016]. We provide their technique with the sparse set of correspondences, which contain facial and peripheral features, as detailed in Section 4. As the figure illustrates, our confidence-aware warping strategy yields visibly-smoother results.

of Thies et al. [2016] extended their work by constructing a mouth database from the *source* sequence to blend it into the target. Even with this extension, as illustrated in Figure 7, they *only* transfer the expression of the face model. Thus, the head of target face is not animated and remains still in an unnatural way and not brought to life as with our technique. Furthermore, the expression (e.g., the smile in Figure 7) generated using our technique is clearly more natural, due to the fine-scale details which are added on top of the image-based warp.

Garrido et al. [2014]: Unlike our approach which animates the full target image, they only reenact the facial region. Their technique resembles ours in the sense that they also employ a 2D warping strategy and avoid mapping the face onto a 3D model. As demonstrated in Figure 7, their technique registers the target face onto the driving video frames, hence the identity of the target is not preserved, even within the facial region.

7.3 Limitations

There are some notable limitations to our work. As demonstrated in Figure 8, we cannot make significant changes to the head pose as we only have the visual information in the single target image. Furthermore, we assume the target image contains a neutral face, and when that assumption is violated, the output frames may seem unnatural. We are also limited by the accuracy of the face tracker we are using. Currently, our results exhibit some temporal artifacts, especially for non-uniform backgrounds. Moreover, there is not enough detail to fully close the eyes of the target image. However, eye-blinking happens quickly in a playing video, and therefore this unnatural affect is almost unnoticeable. A plausible solution should

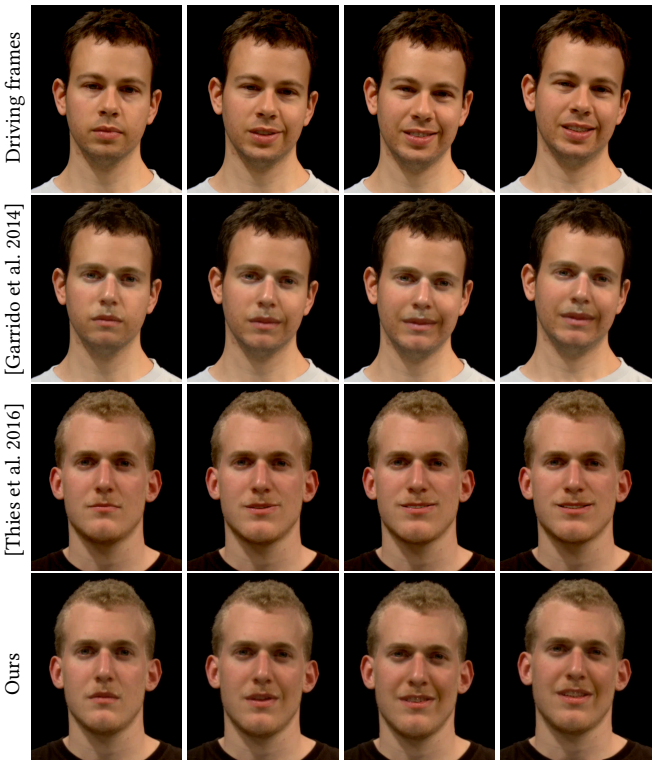


Fig. 7. Comparison to previous reenactment works by replicating the single target image (on the bottom left) to form a video. Representative driving frames appear on top, and the corresponding animated frames below. As the figure illustrates, Garrido et al. [2014] only transfer the facial region and the identity of the target is not preserved, even within this region. In Thies et al. [2016], the head cannot move, hence the generated expressions are significantly less natural-appearing. The full sequences are included in the supplementary material.

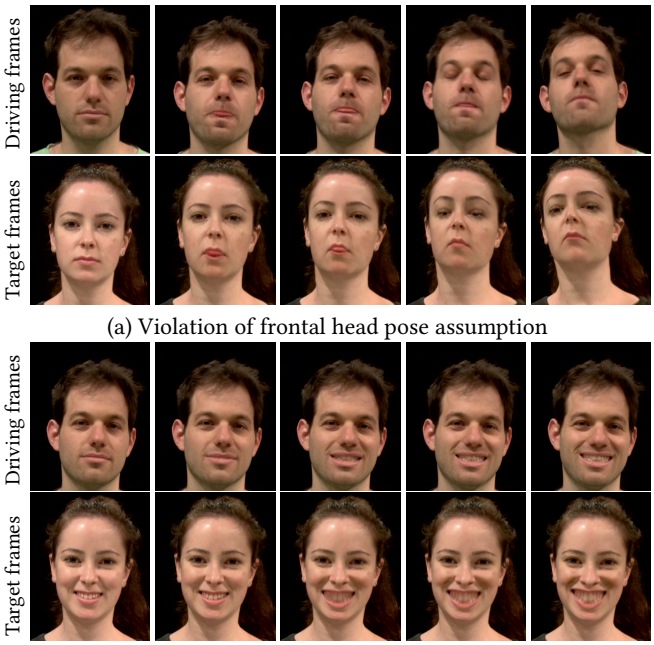
incorporate a personalized "eye closing" manipulation, and would require special treatment of such features, like the "eye opener" of Shu et al. [2016].

7.4 Runtime

Our method is implemented in C++. Given a target image of 600-by-800-pixel resolution, reenacting a driving video of 100 frames (roughly three seconds) takes about 42 seconds on average on a 2.8 GHz Intel Core i7 Mac-Book Pro. For each driving video, we precompute the tracked features and the ratio images for each frame. At runtime, we first warp the target image, the driving video frame, and the ratio image according to our warping procedure. Typical runtime is around 28 seconds for the warp computations. In the remaining 14 seconds, we add the fine-scale details and hallucinate missing regions.

8 APPLICATIONS

Our technique enables automatically animating a still portrait, making it possible for the subject in the static target image to express various emotions. Our primary application is in creating reactive profiles (Section 8.1). We further show that our approach can be directly applied to old portraits and facial avatars (Section 8.2).



(a) Violation of frontal head pose assumption

(b) Violation of neutral target face assumption

Fig. 8. Limitations of our method. (a) Starting from a neutral target image (on the bottom left), our image-based approach eventually breaks as the local geometric features or the head pose significantly change. (b) When the target image violates the neutral target face assumption (on the bottom left), noticeable distortions may occur as the expression evolves. The corresponding driving frames are illustrated on top in both (a) and (b).

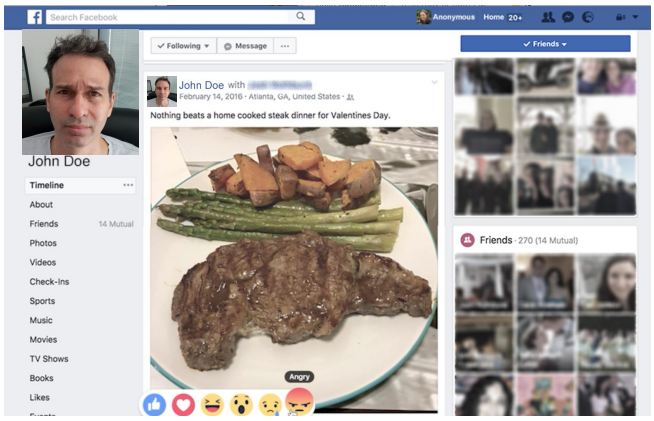


Fig. 9. We generated reactive profiles in the context of a mocked up Facebook page. See the accompanying video for a full demonstration.

8.1 Reactive profiles

Reactive profiles are portraits that can react to different stimuli, or triggers. These moving portraits are probably most well-known from Harry Potter's magical world, where people in photographs

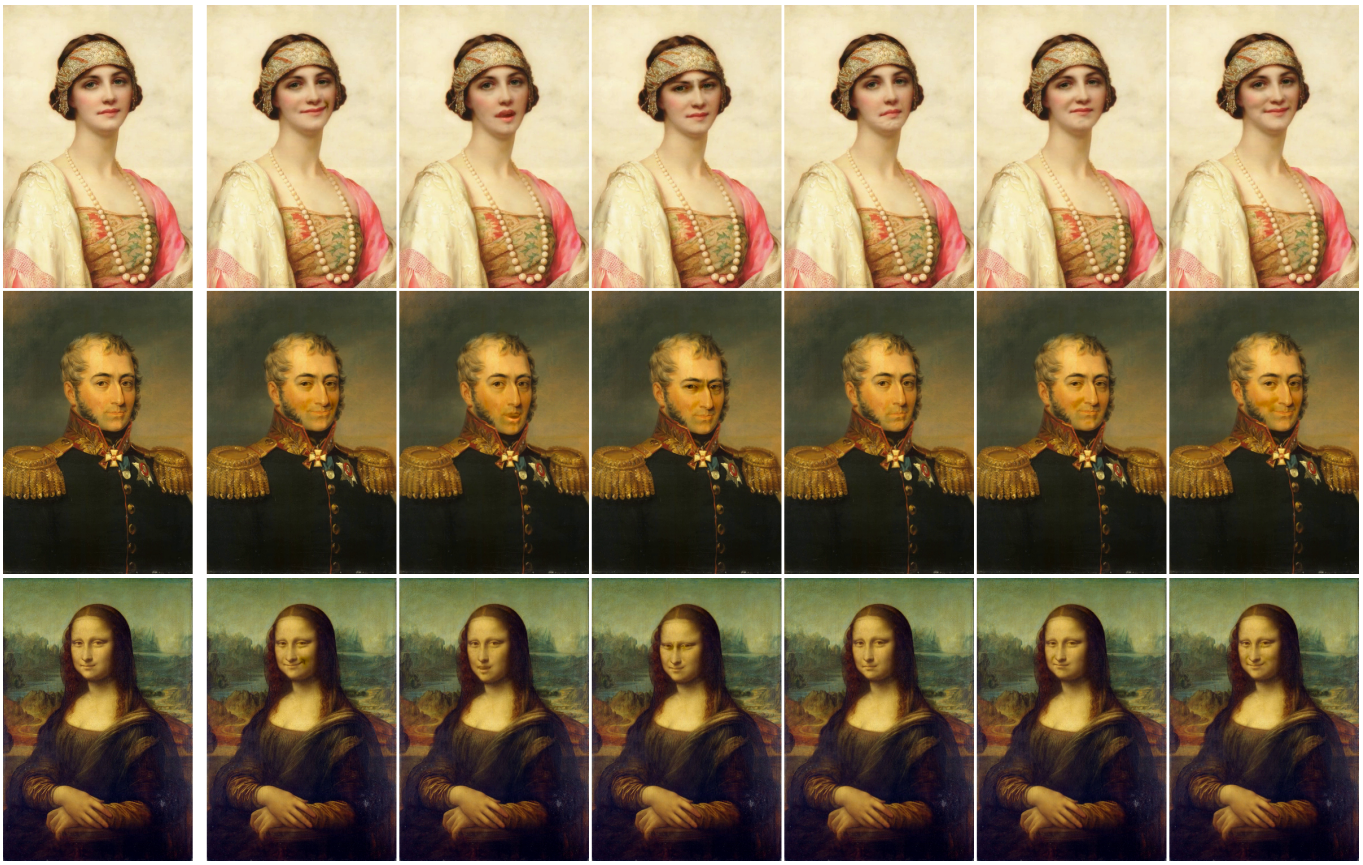


Fig. 10. Animating portraits. The original portraits appear on the left, alongside a few representative frames. Please refer to the accompanying video for the animation clips. ©Wikimedia

move, wave, etc. Our technique enables to automatically generate such moving portraits from a single reference image.

To do so, we used the available driving videos to generate the various emotion animations. We also captured multiple subjects who were asked to stare idly at the camera for a few seconds. Snippets from the "Talking Face Video" [Cootes Cootes] were also used. These additional videos were used to create an *idle* state, to make it feel as if the subject in the portrait is simply breathing, without making any significant facial animations.

We can then transition between the different states, or emotions, on the fly. Using our technique, single target images can follow these driving videos to create new reactive profiles. We created a demo with reactive profiles that react to triggers in real-time. See the accompanying video to view a reactive profile that was generated in the context of a mocked up Facebook page (see Figure 9 for an illustration).

8.2 Non photo-realistic faces

To demonstrate our method on wider scope of facial images, we applied our technique directly on painted portraits and facial avatars (e.g., emojis).

Figure 10 illustrates representative frames (driven according to the driving frames that appear in the first six columns of Figure 13) on a few famous portraits. We are not aware of any previous works that manipulate and animate these portraits completely automatically, as we do. Furthermore, we provide the full animations in the accompanying video.

Taigman et al. [2016] recently presented a technique which enables automatically generating a neutral facial avatar (emoji) from a facial image. Our technique is able to animate these emojis as well to express various emotions. Figure 11 illustrates representative frames (driven according to the driving frames that appear in the first four columns of Figure 13). In the accompanying video, we provide the full animations which were generated automatically using our technique.

As previously stated, we applied our technique *as is*, without adjusting to the painted portrait or emoji domain. To avoid mapping realistic teeth to the emoji animation, we demonstrate this application on a few driving videos where the teeth are not apparent throughout the animation.

Fig. 11. Animating facial avatars (emojis). Emojis (first column) are generated automatically on sample images from the CelebA dataset [Taigman et al. 2016]. Using our technique, these emojis can animate various expressions. Above, we provide representative frames.

9 CONCLUSIONS AND FUTURE WORK

We have presented a method that brings life to a single still portrait in the sense that the image is animated to imitate the facial expressions of a driving video. The use of a single input image captured in casual settings is particularly challenging, but it offers a wide applicability of the method. Technically, the main challenge was to preserve the identity of the target face, while manipulating it with warps and features taken and transferred from frames of some arbitrary driving video. We built on the fact that there is a significant commonality in the way humans "warp" their faces to make an expression. Thus, transferring local warps between aligned faces succeeds in hallucinating facial expressions. We transfer both geometric warps, consisting of 2D offsets, and photometric changes, consisting of illumination ratios.

As we have shown, the internal features of the mouth of the driving video are transplanted to the target face to compensate for the disoccluded region. Although this transplanting violates our goal of keeping the identity of the target face features, the affect of this violation is only secondary, since humans in general are not as sensitive to teeth as a recognition cue. For future work, we plan to consider automatically selecting driving videos that best match the target face to begin with. In this work, we deliberately avoided using such a selection to show the robustness of the method to an arbitrarily selected video.

In the future, we will consider combining our technique with 3D methods. For example, if the face departs from a frontal facing pose, we can map the face over a template 3D face and use the rotated 3D model to allow a wider motion of the facial region. For such non-frontal poses, we can tolerate a certain drop in the quality caused by the re-projection of the image.

We expect other fun applications for the work we have shown here. One can imagine coupling this work with an AI to create an interactive avatar starting from a single photograph.

REFERENCES

Jiamin Bai, Aseem Agarwala, Maneesh Agrawala, and Ravi Ramamoorthi. 2013. Automatic cinemagraph portraits. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 17–25.

Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. 2003. Reanimating faces in images and video. In *Computer graphics forum*, Vol. 22. Wiley Online Library, 641–650.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 187–194.

Jean-Yves Bouguet. 2001. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation* 5, 1-10 (2001), 4.

Pia Breuer, Kwang-In Kim, Wolf Kienzle, Bernhard Scholkopf, and Volker Blanz. 2008. Automatic 3D face reconstruction from single images or video. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 1–8.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 46.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2014), 413–425.

Fig. 12. The intermediate stages of our algorithm: the target image (input), coarse target frame (coarse), transferring hidden regions (+hidden), and our final result, after adding fine-scale details (+fine details), e.g., note the added creases to the right of the nose of the leftmost image. ©Unsplash photographers Brooke Cagle, Ryan Holloway, Beto Silvestre, Drew Graham, William Stitt, Lauren Ferstl, Ryan Holloway, Christopher Campbell, Tanja Heffner and Alex Sheldon.

Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 126.

Erika Chuang and Christoph Bregler. 2005. Mood swings: expressive speech animation. *ACM Transactions on Graphics (TOG)* 24, 2 (2005), 331–347.

T. F. Cootes. Talking face video. *http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html* (????).

Kevin Dale, Kalyan Sunkavalli, Micah K Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. 2011. Video face replacement. *ACM Transactions on Graphics (TOG)* 30, 6 (2011), 130.

Changxing Ding and Dacheng Tao. 2016. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 3 (2016), 37.

Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. 2016. Perspective-aware Manipulation of Portrait Photos. (2016).

Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. 2016. DeepWarp: Photorealistic image resynthesis for gaze manipulation. In *European Conference on Computer Vision*. Springer, 311–326.

Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. 2014. Automatic face reenactment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4217–4224.

Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. 2015. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 193–204.

Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. 2015. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4295–4304.

Alexander Hornung, Ellen Dekkers, and Leif Kobbelt. 2007. Character animation from 2D pictures and 3D motion data. *ACM Transactions on Graphics (TOG)* 26, 1 (2007).

Masahide Kawai, Tomoyori Iwao, Daisuke Mima, Akinobu Maejima, and Shigeo Morishima. 2013. Photorealistic inner mouth expression in speech animation. In *ACM SIGGRAPH 2013 Posters*. ACM, 9.

Masahide Kawai, Tomoyori Iwao, Daisuke Mima, Akinobu Maejima, and Shigeo Morishima. 2014. Data-driven speech animation synthesis focusing on realistic inside of the mouth. *Journal of information processing* 22, 2 (2014), 401–409.

Ira Kemelmacher-Shlizerman, Aditya Sankar, Eli Shechtman, and Steven M Seitz. 2010. Being john malkovich. In *European Conference on Computer Vision*. 341–353.

Davis E King. 2009. Dlib-ml: A machine learning toolkit. *J. Mach. Learning Research* 10 (2009), 1755–1758.

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. 2016. Fast face-swap using convolutional neural networks. *arXiv preprint arXiv:1611.09577* (2016).

Claudia Kuster, Tiberiu Popa, Jean-Charles Bazin, Craig Gotsman, and Markus Gross. 2012. Gaze correction for home video conferencing. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 174.

Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. 2008. Data-driven enhancement of facial attractiveness. In *ACM Transactions on Graphics (TOG)*, Vol. 27. ACM, 38.

Kai Li, Feng Xu, Jue Wang, Qionghai Dai, and Yebin Liu. 2012. A data-driven approach for facial expression synthesis in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 57–64.

Zicheng Liu, Ying Shan, and Zhengyou Zhang. 2001. Expressive expression mapping with ratio images. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 271–276.

Iacopo Masi, Anh Tuan Tran, Jatuporn Toy Leksut, Tal Hassner, and Gérard G. Medioni. 2016. Do We Really Need to Collect Millions of Faces for Effective Face Recognition? *CoRR* abs/1603.07057 (2016). http://arxiv.org/abs/1603.07057

Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*. IEEE, 5–pp.

Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson image editing. In *ACM Transactions on Graphics (TOG)*, Vol. 22. ACM, 313–318.

Marcel Piotraschke and Volker Blanz. 2016. Automated 3d face reconstruction from multiple images using quality measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3418–3427.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, Vol. 23. ACM, 309–314.

Shunsuke Saito, Tianye Li, and Hao Li. 2016. Real-time facial segmentation and performance capture from rgb input. In *European Conference on Computer Vision*. Springer, 244–261.

Fig. 13. Various driving videos animating casual still images. Representative frames are provided above. Please see the accompanying video for animations. ©Unsplash photographers Christopher Campbell, Joseph Gonzalez, Nathan Dumlao and Jimmy Bay.

Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. 2011. Real-time avatar animation from a single image. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 117–124.

Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. 2016. Automatic Portrait Segmentation for Image Stylization. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 93–102.

Zhixin Shu, Eli Shechtman, Dimitris Samaras, and Sunil Hadap. 2016. EyeOpener: Editing Eyes in the Wild. *ACM Transactions on Graphics (TOG)* 36, 1 (2016), 1.

Yaniv Taigman, Adam Polyak, and Lior Wolf. 2016. Unsupervised Cross-Domain Image Generation. *arXiv preprint arXiv:1611.02200* (2016).

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE* 1 (2016).

Michel Valstar and Maja Pantic. 2010. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*. 65.

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face transfer with multilinear models. In *ACM Transactions on Graphics (TOG)*, Vol. 24. ACM, 426–433.

Fei Yang, Lubomir Bourdev, Eli Shechtman, Jue Wang, and Dimitris Metaxas. 2012. Facial expression editing in video using a temporally-smooth factorization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 861–868.

Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. 2011. Expression flow for 3D-aware face component transfer. In *ACM Transactions on Graphics (TOG)*, Vol. 30. ACM, 60.

Raymond Yeh, Ziwei Liu, Dan B Goldman, and Aseem Agarwala. 2016. Semantic Facial Expression Editing using Autoencoded Flow. *arXiv preprint arXiv:1611.09961* (2016).

Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. 2010. Parametric reshaping of human bodies in images. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 126.