Privacy-Preserving Annotation of Face Images through Attribute-Preserving Face Synthesis

Sola Shirai
Worcester Polytechnic Institute
Worcester, MA
sshirai@wpi.edu

Jacob Whitehill
Worcester Polytechnic Institute
Worcester, MA
jrwhitehill@wpi.edu

Abstract

We investigate the viability of collecting annotations for face images while preserving privacy by using synthesized images as surrogates. We compare two approaches: a state-of-the-art 3-D face model based on deep neural networks (Extreme3D [24]) to render a detailed 3-D reconstruction of the face from an input image; and a novel generative adversarial network architecture that we propose that extends BEGAN-CS [5] to generate images conditioned on desired low-level facial attributes. Using these two alternative models, we conduct experiments on Mechanical Turk to annotate emotions ("joy" and "anger") on raw and synthesized versions of face images. Across 60 workers each annotating 3 versions of 60 images in each experiment, we find that: (1) The labeling accuracy when viewing surrogate images can be very similar to the accuracy when viewing raw images, but depends significantly on the labeling task. (2) The proposed extension to BEGAN-CS is effective in generating realistic images that correspond to the input vector of low-level facial attributes. (3) Overall, the GAN-based approach to generating surrogate images gives comparable accuracy as the 3-D face model, but is easier to train.

1. Introduction

Computer vision for automatic face analysis often requires large-scale annotation of face images to provide labeled training and testing data. While crowdsourcing (e.g., on Mechanical Turk) can help to reduce the time and expense of annotation, it is not always practical for *privacy* reasons: in many educational [20], medical [23], or military settings, the face data that are collected might contain images that are sensitive and show people in distressing or potentially embarrassing situations, and it would be unacceptable to post such images on the Web for public viewing. For instance, in order to train an automatic facial expression recognition system to recognize *pain*, one might collaborate with a hospital to collect images of patients with medical conditions that are known to be painful, and then train a classifier designed to estimate the intensity of pain. In another scenario, an educational researcher might want to create an automatic student "engagement" detector and would thus record videos of children in school classrooms. In both these scenarios, the face images should likely be kept private to only the research team.

One strategy to facilitate efficient labeling while maintaining data privacy – which has been theoretically possible for some time but rarely practiced – is to perform some form of *face de-identification* [11] of the images while retaining enough information for human workers to be able to accurately assign labels. Naive face de-identification methods such as blurring or distorting the image tend to remove too much information for human labelers to label accurately. However, deep learning provides new ways of preserving the important facial attributes while removing identity information. In this paper we explore a de-identification approach based on *automatic image synthesis* whereby either a Generative Adversarial Network (GANs) [10], or a 3-D generative face model that uses a neural network to estimate shape parameters [14], is used to render high-resolution, realistic face images. Importantly, our approach synthesizes new images that can be *conditioned* on specific attributes such as gender, age, head pose, facial expression, etc. Most existing approaches to face de-identification offer k-anonymity guarantees [22] but may still contain some information about the raw image. In contrast, the synthesized images in our method contain *zero* information about the original faces (other than the attributes themselves that are to be labeled).

We focus on computer vision tasks in which the goal is to train a classifier of a *high-level* facial state – e.g., pain, anxiety, engagement – that may be expressed by *low-level* facial features. For example, human facial expressions can be characterized precisely by the intensity of 40+ different facial *action units* (AUs) that correspond approximately to different muscle groups ([9]). While basic emotions such as

"joy" and "anger" are sometimes conceptualized as primitive states, their facial expression can be decomposed into the presence or absence of different AU combinations, e.g., AUs 6+12 for joy, and AUs 4+5+7+23 for anger.

The basic idea we propose is the following (see Figure 1): Suppose we want to train a face classifier to estimate the intensity of a high-level state such as student engagement. We first collect a large dataset of face images from students. Then, from each image in our dataset, we use a pre-trained, off-the-shelf automatic facial expression recognition tool (e.g., OpenFace [1]) to extract low-level facial attributes such as head pose, eye closure, facial action units, etc. (These features were shown to be related to perceived student engagement in [28].) These low-level attributes are then fed to a GAN or 3-D face model to produce a new *surrogate image* that resembles the raw input image in many ways *except* facial identity. Then, the set of all surrogate images is posted to a crowdsourcing site such as Mechanical Turk for annotation. Assuming the surrogate images preserve the relevant attributes (the intensity of "engagement" in our example), their labels can then be assigned to the raw images and used to train an automatic student engagement classifier.

Contributions: (1) We propose an approach to privacy-preserving face image annotation based on synthesizing surrogate images that contain zero information about the original faces other than a specifiable set of attributes. (2) We propose a novel enhancement to the BEGAN-CS [5] architecture whereby images can be conditioned on specified facial attributes (e.g., gender, head pose, AUs). (3) We conduct an experiment on Mechanical Turk ($N = 60$ participants) to compare the annotation accuracy (for the emotional states of "joy" and "anger") of raw images, surrogate images generated from a GAN, and surrogate images generated from a state-of-the-art 3-D face model. Our results suggest that, while there is an accuracy loss when labeling surrogate images, the approach is promising and will likely improve with better trained GANs.

2. Related Work

2.1. Face De-Identification

Face de-identification algorithms [12, 15, 18, 29, 16, 2] remove identity information from images. This can be achieved using simplistic methods such as applying significant blurs, pixelization, or black boxes on faces. However, this leads to a significant loss of facial information, making it impractical for our setting. Other approaches use the k-same algorithm [18] to combine k different images of faces that are very similar. This results in an image that is mostly de-identified while retaining important low-level features; however in practice this approach often still leaves significant facial artifacts [12].

GANs for face de-identification: Similar to our work, several researchers have also explored using GANs to de-identify faces. Works such as [8, 19, 29] utilize auto-encoder based GANs to modify input images with the objective of minimizing identity information while retaining underlying features (*e.g.* structure, expression). [2, 16] show GAN approaches to disentangle identity and facial attributes of images, allowing the generator to synthesize new faces using only the extracted attributes of an input face. [21] demonstrates a method of generating images with less direct input from the raw images, where facial landmarks are used by the generator to synthesize faces. Our paper primarily differs from these works in that our model uses no information from the raw images (other than a specifiable set of low-level attributes) to generate faces, and our evaluation's focus is on collecting high-level annotations on synthesized images.

2.2. 3-D Face Models

Techniques for face image synthesis such as 3-D morphable models (3DMM) [4] work by transforming and fitting some base 3-D face model to a desired target shape. In this paper, we look in particular at approaches that use convolutional neural networks (CNNs) to fit 3-D face models. One notable example is 3-D Dense Face Alignment (3DDFA) [31], which fits and aligns a 3DMM to the input image using cascading CNNs. Another is Expression Net (ExpNet) [7], which performs regression directly on 3DMM expression coefficients rather than detecting and using facial landmarks.

For our crowdsourcing experiments involving the use of 3-D face models, we chose Extreme3D [24] which is publicly available online [1]. Extreme3D combines strong regularization for the overall face shape with weak regularization for more local details of the face. The Extreme3D model separately computes a foundation shape, facial expression, and viewpoint of a face. It then estimates a bump map to capture mid-level features. Finally, occluded details of the face are added on to produce the final output 3-D face.

3. Proposed Solution

To collect annotations for face images while preserving privacy, we propose the approach of generating new surrogate images that preserve the essential facial attributes that we wish to label (Figure 1 illustrates the general workflow). For face image synthesis, we consider two alternative modern approaches, both of which retain 0 information about the raw images (except the attributes such as facial expression, gender, and head pose): (1) We can use a 3-D face model such as Extreme3D [24] that takes a face image as input, estimate 3-D face model parameters using a neural

[1] https://github.com/anhttran/extreme_3d_faces

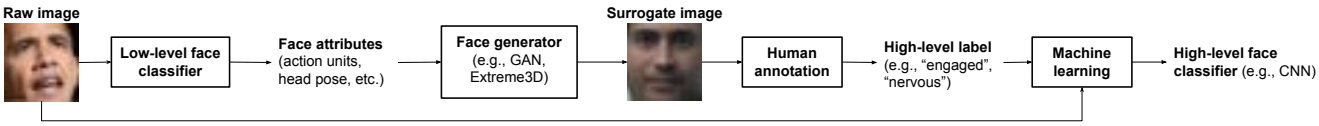


Figure 1. Pipeline for proposed solution to produce surrogate face image to collect high-level labels.

network, and then be used to synthesize a 2-D face image using these parameters. Extreme3D uses deep neural networks to detect how to modify the expressions and shapes of the base 3-D face model, as well as what pose to position the face in. It can also capture finer details in the face such as wrinkles. (2) We can train a Conditional GANs [17] that takes both a noise vector and specific face attributes as input to generate surrogate face images that contain the same low-level features (*e.g.* expression, gender, pose) as the raw images.

4. Conditional BEGAN-CS Architecture

We use BEGAN-CS as the backbone of our GAN-based approach because, in pilot experimentation, it resulted in considerably less mode collapse in generated images than the regular GAN [10]. We propose a novel enhancement to the BEGAN-CS [5] architecture that enables the network to synthesize new images from both a noise vector as well as a specified vector of low-level facial attributes (head pose, gender, and AUs). BEGAN-CS [5] is based on the BEGAN by [3], which in turn is based on EBGAN by [30]. In all these networks, the discriminator is an auto-encoder that is trained on the reconstruction error of input images (minimizing error for real images and maximizing it for fake images). BEGAN expands upon EBGAN by introducing an equilibrium enforcing term to balance the weighting of reconstructing real and generated images, which can result in more stable training. BEGAN-CS expands upon BEGAN by adding a constraint on the internal state of the auto-encoder, limiting the difference between it and the input noise for generated images. The addition of this constraint is shown to improve mode collapse in the BEGAN-CS model. We further build upon BEGAN-CS's architecture to introduce conditional training.

4.1. BEGAN-CS Overview

Let $D : R^{N_x} \mapsto R^{N_x}$ be the discriminator (an auto-encoder) and $G : R^{N_z} \mapsto R^{N_x}$ be the generator, where N_x is the dimension of the real data x and N_z is the dimension of our noise input z. G takes a random noise vector z_G as input and produces an image $G(z_G; \theta_G)$ using parameters θ_G. The discriminator is an auto-encoder; the encoder part Enc produces a latent code z_D. Finally, let $\mathcal{L}(v) = |v - D(v)|^2$ for $v \in R^{N_x}$ compute a reconstruction loss.

In the original BEGAN-CS, there are three loss functions: $\mathcal{L}_G$ for the generator, $\mathcal{L}_D$ for the discriminator, and also a *constraint loss* $\mathcal{L}_C$:

$$\begin{aligned}
\mathcal{L}_D &= \mathcal{L}(x;\theta_D) - k_t \cdot \mathcal{L}(G(z_G;\theta_G);\theta_D) + \alpha \cdot \mathcal{L}_C \\
\mathcal{L}_G &= \mathcal{L}(G(z_G;\theta_G);\theta_D) \\
\mathcal{L}_C &= \|z_D - Enc(G(z_D))\| \\
k_{t+1} &= k_t + \lambda(\gamma\mathcal{L}(x;\theta_D) - \mathcal{L}(G(z_G;\theta_G);\theta_D))
\end{aligned}$$

Here, x is a real image, and $G(z_G; \theta_G)$ is a generated (fake) image. The term k_t helps to stabilize the training process by maintaining a balance between the reconstruction loss of real and generated data. BEGAN-CS introduces the *constraint loss* $\mathcal{L}_C$, which enforces that the internal state of the encoder for generated data, $Enc(G(z))$, resembles the input noise into the generator. γ and α are hyperparameters.

4.2. Conditional Image Generation

To enable conditional training on labels, we introduce an auxiliary predictor for labels into the BEGAN-CS discriminator network. This auxiliary network is a simple fully connected network that takes the internal state of the encoder as input and outputs predictions of labels. Additionally, in the generator, we concatenate the label information with the input noise.

To accommodate this new predictor, the loss functions for the generator and discriminator have a new loss added. Since the labels for our dataset were continuous values, we chose to use mean squared error (MSE) to measure the error of our predictor. Given input ground-truth labels y and predicted labels $\hat{y}$, the loss functions are then updated with an MSE loss term $\mathcal{L}_{MSE}$.

$$\begin{aligned}
\mathcal{L}_D &= \mathcal{L}(x;\theta_D) - k_t \cdot \mathcal{L}(G(z_G|y;\theta_G);\theta_D) + \\
&\quad \alpha \cdot \mathcal{L}_C + \mathcal{L}_{MSE}(y;\hat{y}) \\
\mathcal{L}_G &= \mathcal{L}(G(z_G|y;\theta_G);\theta_D) + \mathcal{L}_{MSE}(y;\hat{y})
\end{aligned}$$

Figures 2 and 3 – in particular the label vector y and auxiliary predictor network – illustrate our proposed enhancement to BEGAN-CS. The generator concatenates a vector of random noise z and labels y as input to generate an image. The discriminator takes either a generated image x_{fake} or real image x_{real} as input, encodes the image into a vector $\hat{z}$, and produces a reconstructed image $D(x)$ and a predicted label $\hat{y}$.

In the discriminator's loss function, we only use the MSE on label predictions over real images to allow the auxiliary

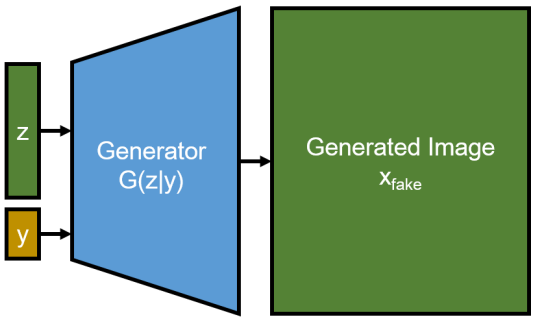


Figure 2. Overview of the final generator network.

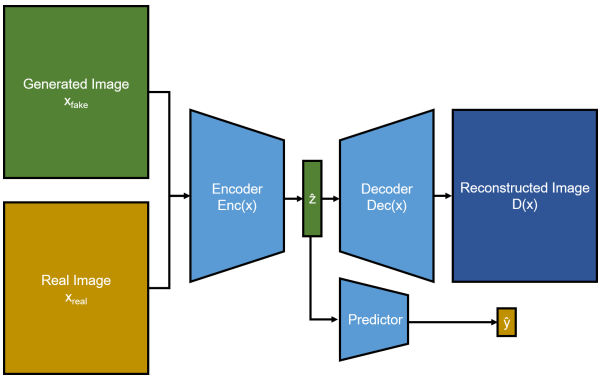


Figure 3. Overview of the final discriminator network. Note that only one of the inputs x_{real} and x_{fake} are passed through the network at a time.

network to learn to more accurately predict labels from the internal state of the auto-encoder. The generator's loss function is amended to include the MSE on labels predicted by the auxiliary network on the generated images. Intuitively, our discriminator's training objectives are 1) minimize reconstruction error for real images, 2) maximize reconstruction error for fake images, 3) minimize difference between generator input z and the encoding $\hat{z}$ of the generated image, and 4) minimize the difference between the real label y and predicted label $\hat{y}$ for real images. This setup encourages the generator to produce images that reflect the input labels in a similar fashion to traditional conditional GANs [17]; even if the generator is producing extremely realistic faces, the discriminator will learn to penalize fake images because the predicted labels will not match the ground truth.

Although it was not used for our final experiments, we also explored an alternate model design for the discriminator without the auxiliary network to predict labels. Instead, this alternate design used an encoder that produced a latent vector of size $|z| + |y|$. This vector was then split into two vectors (to act as $\hat{z}$ and $\hat{y}$) to calculate $\mathcal{L}_{MSE}$ and $\mathcal{L}_C$. Intuitively this design aimed to train the discriminator to encode generated images directly into the input noise and label used to generate the images, which we supposed would in turn help train the generator to create images that more closely reflected the desired features. However, empirical results of image generation on the alternative design were generally lower quality than those produced by our final model.

4.3. Model Hyper-parameters

We trained our final GAN model on approximately 18000 images from the GENKI1M dataset [25] (each image was 64 x 64 grayscale) with 10-dimensional low-level attribute vectors y that were estimated using the Emotient SDK facial expression recognition software [13]: yaw, pitch, roll, the intensity of 5 facial action units (AU 4, 5, 6, 7, 12, and 23), and gender (probability of male face). As a preprocessing step, these real-valued labels were normalized to values between 0 and 1.

Input into the generator network was a 64 dimensional noise vector, sampled uniformly between -1 and 1. Concatenating this with input labels, the input was passed through convolutional layers and upsampled to form the 64 x 64 image outputs. The Adam optimizer was used to train the discriminator, with a learning rate of 0.0001.

In the discriminator network, we used convolutional layers in the encoder to encode the input 64 x 64 image into a 64 dimensional vector. The decoder portion of the discriminator had an identical shape to the generator network (other than inputting labels). The auxiliary predictor was a 2 layer fully connected network, taking the image encoding as input and producing label predictions. We set the values of α and γ to 0.5 and 0.1, respectively, and we once again used the Adam optimizer with a learning rate of 0.0001 for the training process.

4.4. Generated Image Examples

Figure 4 shows a comparison of images from the GENKI-4K dataset (in the top row), followed by images generated by our GAN using the AU, gender, and pose labels, and a rendering of the 3-D face model generated by Extreme3D. The joy and gender features appear to be reflected fairly well in our examples of GAN images, but pose information is not captured well by the GAN.

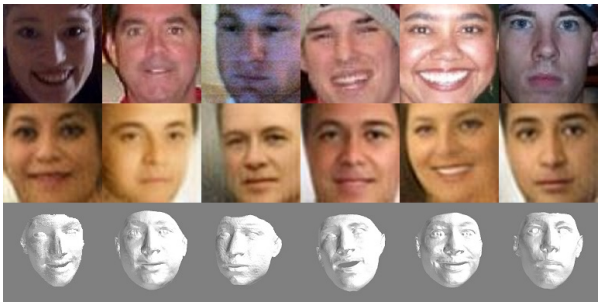


Figure 4. Examples comparing real images to images generated by the GAN and Extreme3D.

5. Experiment

We conducted experiments on Amazon Mechanical Turk to compare the labeling accuracy of surrogate images generated by our GAN and the Extreme3D face model, as well as the raw images. We analyzed how labeling accuracy increased with the number of assigned labelers (where majority vote was used to aggregate votes for each image), as well as the degree of uncertainty in labeling each image (what fraction of the labelers agreed on each image) using each approach. Finally, we examine failure modes of the GAN and Extreme3D models.

As the target labeling tasks, we chose the emotional states of "joy" and "anger". Though often regarded as primitive states, their expression in human faces actually consist of a constellation of multiple facial muscles (AUs). We chose this labeling task because it is simple for labelers to understand and still enables us to assess our approach to privacy-preserving large-scale annotation.

5.1. Data

Images for the experiment were selected from the GENKI-4K [26] dataset, which contains many positive and negative joy emotions as well as some images that contain negative emotions (e.g., anger). We used the Emotient SDK facial analysis software [13] to estimate the degree of "joy" and "anger" in each image. With the goal of sampling images with a variety of hard-to-label and easy-to-label images, we selected images according to Emotient's estimate of the log-likelihood ratio (joy versus not joy, anger verus not anger) of each emotion. *Joy*: We binned the images with a bin width of 0.5 (where the majority of evidence scores lied between -3 and 3). From each bin, we sampled 5 images (for a total of 60 images). *Anger*: Since relatively few GENKI-4K images contained high levels of "anger", we randomly sampled 30 images for which the log-likelihood ratio of anger was negative and 30 images for which it was positive.

For each of the 60 selected images for each emotion labeling task, we generated two surrogate images using the alternative approaches. For the 3-D face version, we generated and rendered 3-D meshes from the raw images using the Extreme3D system. For our GAN-generated versions, we collected low-level feature labels from Emotient for our 60 selected images. These features were passed into the GAN to generate new faces conditioned on the feature labels from the raw image set.

5.2. Experiment Setup

Two experiments (each with $N = 60$ participants) were conducted on Mechanical Turk: one for "joy" labeling, and one for "anger" labeling. All workers were shown a complete set of images – the raw, 3-D face, and GAN-generated versions, for a total of 180 images – for their task. Labeling all images took roughly 10 minutes, and workers were paid $1 for completing the task.

5.3. Evaluation

For each image and each task, we took the majority label of the *raw* images over all 60 labelers as the ground-truth label. We then compared the three image versions (raw, GAN, or Extreme3D (E3D)) for labeling accuracy. For statistical significance testing, we used 1-sample t-tests on the difference in number of labelers who labeled each image correctly between each pair of image types (raw-GAN, raw-E3D, and GAN-E3D).

6. Experimental Results

6.1. Average accuracy of individual labelers

In Table 1, the left-most column shows the average labeling accuracy over individual labelers compared to the ground-truth for each image version and each task.

Accuracy of surrogate images: The results confirm that labels of the surrogate images are highly predictive of the labels that are assigned to the raw images, which provides a basic validation of the proposed labeling pipeline (Figure 1). Moreover, the extension to the BEGAN-CS that are proposed (Section 4.2), whereby images can be conditioned in low-level attributes estimated from off-the-shelf automated facial expression recognition software, were just as good as for a state-of-the-art 3-D face model (Extreme3D). The accuracy difference between the two surrogate image methods (GAN, Extreme3D) was not stat. sig. ($t = 0.48, p = 0.64$ for joy, $t = 1.51, p = 0.14$ for anger).

Labeling surrogate versus raw images: While the surrogate labeling approach shows some promise, the accuracy was still lower than when labeling the raw images: For joy, the accuracy loss is about 10% both for the GAN images the Extreme3D images; both differences were stat. sig. ($t = 4.37, p < 0.001$ and $t = 5.50, p < 0.001$, respectively). However, for anger, the drop in labeling accuracy, compared to labeling the raw images, of either the GAN or Extreme3D surrogates was less severe. In fact, the difference between raw images and GAN was less than 5% and the difference was not stat. sig. ($t = 1.31, p = 0.20$).

6.2. Accuracy for images with $\geq X\%$ consensus

In Table 1, the vote (X%) columns show the labeling accuracy on those images for which at least X% of the labelers agreed on the label. Not surprisingly, by restricting our labeling process to those images with high agreement, we can ensure higher accuracy of the labels. (Note that, since the majority vote of each raw image was taken as the ground-truth, the accuracies for these columns would be 100% by definition and are not displayed.)

Type	Avg	Vote (50%)	Vote (75%)	Vote (90%)
Accuracy: Joy Labeling Task				
RAW	89.0	-	-	-
GAN	71.2	76.7	82.2	84.4
E3D	73.5	81.7	83.7	90.0
Accuracy: Anger Labeling Task				
RAW	79.9	-	-	-
GAN	75.5	76.7	88.4	96.9
E3D	71.6	75.0	78.7	96.5

Table 1. Accuracy of worker labels provided for each type of image: raw images (RAW), generated by GAN (GAN), and rendered by Extreme3D (E3D). Vote (%) gives majority vote accuracy for images where over X% of workers gave the same label.

When we take the majority vote ($\geq 50\%$), both the GAN and Extreme3D versions show improvement in accuracy. Moreover, if we require even higher consensus (75%, 90%), the accuracy for anger labeling approaches 100%. For joy, however, the accuracy is still substantially below 100%. Naturally, only a subset of all images have such a high degree of consensus

6.3. Label ambiguity of surrogate images

Figure 5 shows a histogram (over the 60 images for each image version and each labeling task) of the fraction of labelers who agreed with each other (0.5 to 1.0) when labeling each image. For joy, the raw images show a much higher degree of consensus (greater mass around 1.0), suggesting that the surrogate images (for both GAN and Extreme3D) are often hard to discern (ambiguous) in terms of "joy". However, for anger, the GAN surrogate images behave much better, with label consensus at least as high as for raw images.

6.4. Majority vote over n labelers

Figure 6 shows the labeling accuracy when computing the majority vote across a subset of n labelers for each image, when n is varied from 3 to 60. Each point in each line plot is computed by sampling n labels from our dataset of 60 total labelers. The results suggest that accuracy of the surrogate images can be increased by around 5-10% by sampling labels for each image from about 10-15 labelers, but larger n make little difference. (Note that the curve for the raw images approaches 1.0 by definition of how we computed ground-truth.)

6.5. Accuracy vs Emotion Intensity

We computed the labeling accuracy, for each image version, as a function of the *intensity* of each emotion, as estimated using the output of the Emotient SDK software. Though we used only low-level features (AUs, head pose, etc.) when training the GAN and generating the surrogate images, Emotient can also estimate the intensity "joy"

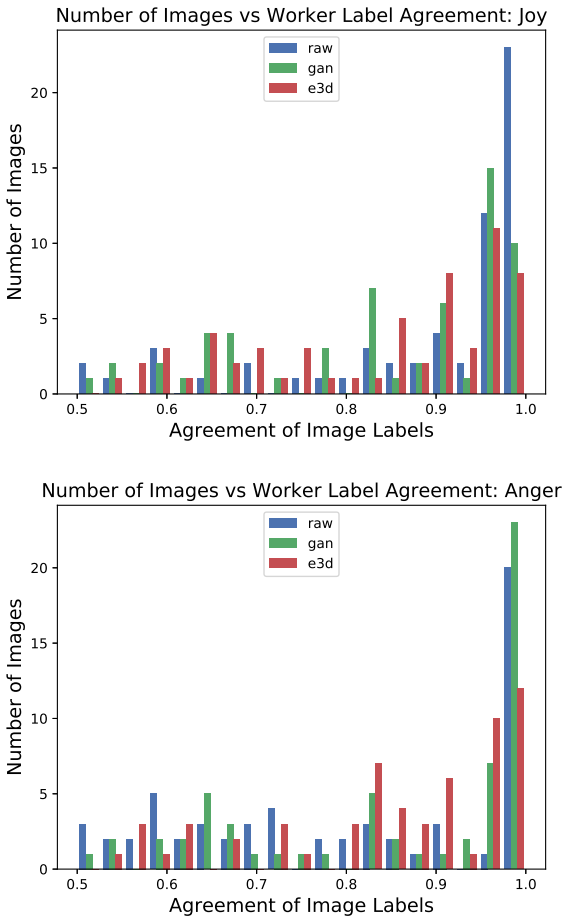
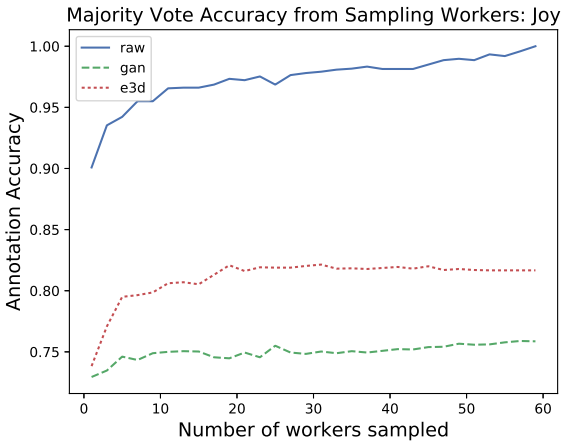


Figure 5. Agreement of image labels (as the proportion of workers who gave the same label) compared to the number of images with that level of label agreement.

and "anger" directly. Though Emotient technically outputs an "evidence" aka log-likelihood ratio rather than intensity, prior work [27] has shown that these are often highly correlated. We then investigated whether the more ambiguous images (i.e., evidence around 0) are more difficult to annotate correctly.

Figure 7 and 8 plot the average annotation accuracy per-image, grouping and averaging image accuracy together based on the joy and anger evidence (as reported by Emotient) of each image. Note that for the raw images, accuracy will always be over 50% since we define the ground-truth based on the majority label given to the raw images. In general, we can see that the GAN images tend to have larger differences in accuracy when the evidence is near 0. While the Extreme3D images also somewhat show a similar pattern of poor performance near 0 emotion evidence, it does not appear to show improvements in accuracy for higher evidences as clearly as the GAN images do (*e.g.* Extreme3D accuracy for positive anger evidence in Figure 8 is decreasing).



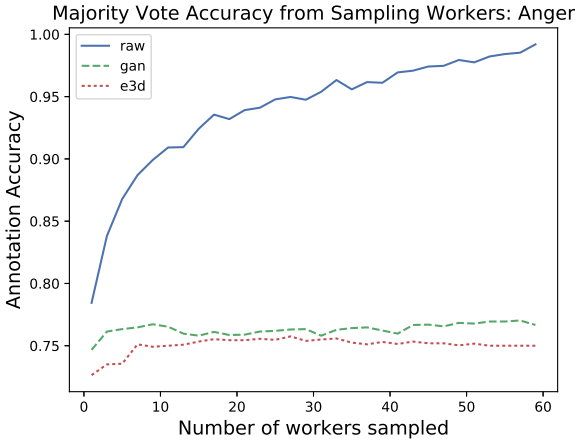


Figure 6. Majority vote accuracy each version of images for the two labeling tasks when sampling a subset of workers.

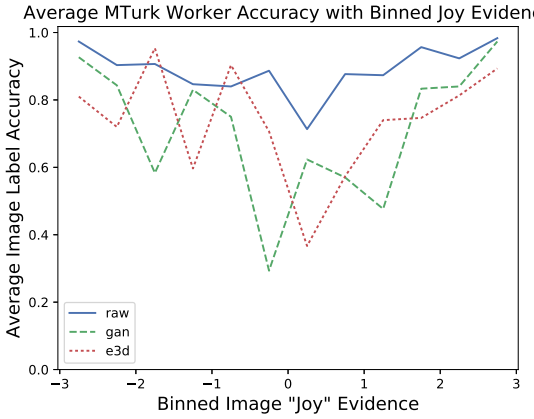


Figure 7. Average worker accuracy for sets of images binned by their joy evidence.

One possible explanation for these results is that the low-level state estimates (AUs, head pose, etc.) themselves are noisy, and this will necessarily degrade performance in the downstream generation of surrogate images. Another is that

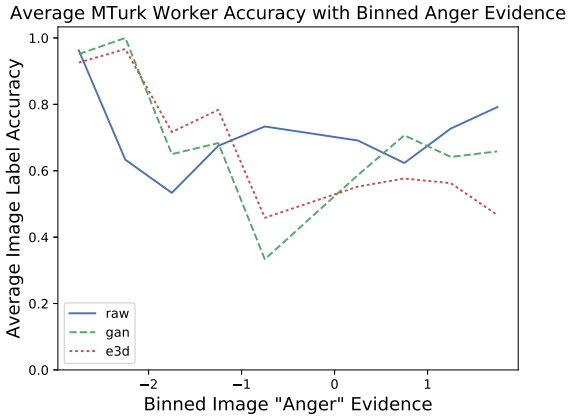


Figure 8. Average worker accuracy for sets of images binned by their anger evidence.

the raw images themselves become harder to label as the intensity of the emotion (anger versus not anger, joy versus not joy) is lower in magnitude.

Accuracy on unambiguous images: When we compute the average accuracy of individual labelers on *ambiguous* images, which we define as those for which the absolute value of Emotient's estimate of "joy" or "anger" evidence is at least 1.0, we obtain slightly better results for all three image versions: For joy, the raw images give an accuracy of 92.1%, the GAN images 78.8%, and the Extreme3D images 78.4% – these numbers are higher than the corresponding numbers in the first column of Table 1, but the gap actually widened. However, for anger, the gap narrows: the accuracy for raw images is 86.8%, for GAN images it is 85%, and for Extreme3D images it is 81.%.

6.6. Surrogate Images with Poor Accuracy

Figure 9 shows 10 pairs of images (raw and GAN versions) that had the lowest labeling accuracy for the GAN versions on the joy labeling task. Labeling accuracy of these GAN images ranged from 33% to 0%. There are several aspects of the raw images shown here that may have contributed to either collecting inaccurate low-level feature labels (which were used to generate the GAN images) or change how workers perceived the emotion displayed by the face. Possible factors include faces with mustaches, large head pose, poor lighting, and mouths that are open. We also see an example of an image generated by our GAN that is extremely blurry, which was found to occur in some instances.

A similar set of 10 pairs of images that had the worst labeling accuracy for Extreme3D versions of images is shown in Figure 10, with accuracy ranging from 43% to 3%. Here we can once again see some instances of images with mustaches, which seem to make the 3-D face renderings become quite bumpy around the mouth. We can also see some ar-

Figure 9. Joy task images with low label accuracy by workers on the GAN versions of the images.

tifacts like strands of hair interfering with the face. The openness of the mouth in some of the 3-D faces appears to not match the raw images, which also likely contributed to more incorrect labels.

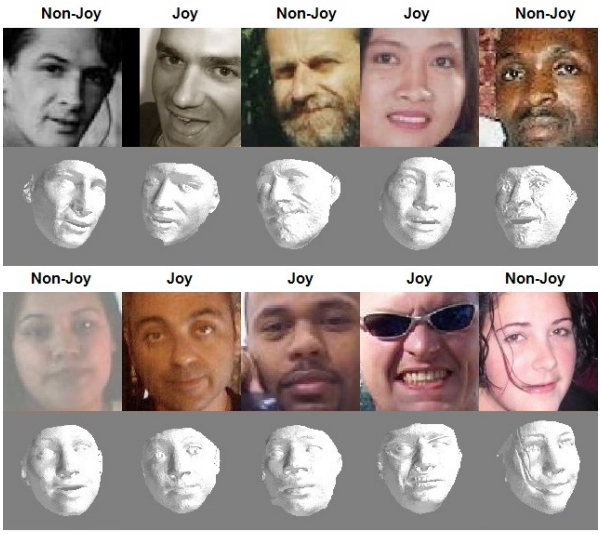


Figure 10. Joy task images with low label accuracy by workers on the Extreme3D versions of the images.

Similar patterns were observed in our anger labeling task for the generated images with poor labeling accuracy. Across the images we tested, we generally see a pattern where images that display less intense joy evidence tended to be harder to label using surrogate images. Comparing the GAN to Extreme3D versions, we see that Extreme3D captures head pose much better.

7. Conclusion

In this paper we explored the viability of using synthesized face images to collect annotations while preserving privacy. In contrast to most face de-identification methods that offer k-anonymity guarantees, our method retains 0 information about the original images, except the low-level attributes (facial action units, gender, pose) themselves. In our experiments to crowdsource expression labels, we see promising results especially for faces that display the expression strongly. On a "joy" labeling task, synthetic images generated by our GAN and Extreme3D an average labeling accuracy of 71.2% and 73.5%, increasing up to 84.4% and 90.0% when only considering images with high label consensus, compared to the 89.0% accuracy of workers on the raw images. The results are even closer for the "anger" labeling task, with labeling accuracy on raw images of 79.9% while GAN and Extreme3D images achieve 75.5% and 71.6%, increasing to 96.9% and 96.5% for high label consensus images.

For the development of our GAN model, we successfully demonstrate a method to introduce conditional image generation into the BEGAN-CS architecture. The addition of our auxiliary prediction network shows to successfully motivate the generator to produce images that align with the desired features used as input, allowing it to generate face images that reflect features from raw images. Furthermore, as the generated images are only exposed to low-level feature information about the raw images, privacy is ensured when using the surrogate images.

Between the GAN and 3-D model (e.g., Extreme3D) approach, we tentatively conclude that the GANs hold more promise: Compared to using Extreme3D, our GAN shows approximately equal performance for our surrogate image labeling approach. Additionally, our GAN is easier to train; compared to the roughly 16,000 images used in our training, Extreme3D uses multiple deep-learning models trained on image sets as large as 2.6 million images [6]. With GANs trained on larger datasets, the veracity of the resulting surrogate images could easily increase.

References

[1] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016. 2

[2] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. pages 6713–6722, 2018. 2

[3] D. Berthelot, T. Schumm, and L. Metz. BEGAN: boundary equilibrium generative adversarial networks. *CoRR*, abs/1703.10717, 2017. 3

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on*

Computer Graphics and Interactive Techniques, SIGGRAPH 1999, Los Angeles, CA, USA, August 8-13, 1999, pages 187–194, 1999. 2

[5] C. Chang, C. H. Lin, C. Lee, D. Juan, W. Wei, and H. Chen. Escaping from collapsing modes in a constrained space. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 212–227, 2018. 1, 2, 3

[6] F. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. G. Medioni. Faceposenet: Making a case for landmark-free face alignment. *CoRR*, abs/1708.07517, 2017. 8

[7] F. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. G. Medioni. Expnet: Landmark-free, deep, 3d facial expressions. *CoRR*, abs/1802.00542, 2018. 2

[8] J. Chen, J. Konrad, and P. Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition. *CoRR*, abs/1803.07100, 2018. 2

[9] P. Ekman and W. V. Friesen. *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978. 1

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. 1, 3

[11] R. Gross, L. Sweeney, J. Cohn, F. De la Torre, and S. Baker. *Face De-identification*, pages 129–146. 07 2009. 1

[12] R. Gross, L. Sweeney, J. Cohn, F. De la Torre, and S. Baker. Face de-identification. In *Protecting privacy in video surveillance*, pages 129–146. Springer, 2009. 2

[13] iMotions. Emotient. https://imotions.com/emotient/. 4, 5

[14] X. Z. Jianzhu Guo and Z. Lei. 3ddfa. https://github.com/cleardusk/3DDFA, 2018. 1

[15] G. Letournel, A. Bugeau, V. T. Ta, and J. P. Domenger. Face de-identification with expressions preservation. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4366–4370, Sep. 2015. 2

[16] Y. Li and S. Lyu. De-identification without losing faces. *arXiv e-prints*, page arXiv:1902.04202, Feb 2019. 2

[17] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 3, 4

[18] E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Trans. on Knowl. and Data Eng.*, 17(2):232–243, Feb. 2005. 2

[19] Z. Ren, Y. J. Lee, and M. S. Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, pages 639–655, 2018. 2

[20] A. Sarrafzadeh, H. Gholamhosseini, C. Fan, and S. Overmeyer. Facial expression analysis for estimating learner's emotional state in intelligent tutoring systems. pages 336–337, 08 2003. 1

[21] Q. Sun, L. Ma, S. J. Oh, L. V. Gool, B. Schiele, and M. Fritz. Natural and effective obfuscation by head inpainting. *CoRR*, abs/1711.09001, 2017. 2

[22] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002. 1

[23] J. Thevenot, M. B. Lpez, and A. Hadid. A survey on computer vision for assistive medical diagnosis from faces. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1497–1511, Sep. 2018. 1

[24] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. G. Medioni. Extreme 3d face reconstruction: Looking past occlusions. *CoRR*, abs/1712.05083, 2017. 1, 2

[25] http://mplab.ucsd.edu. The MPLab GENKI Database. 4

[26] http://mplab.ucsd.edu. The MPLab GENKI Database, GENKI-4K Subset. 5

[27] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):2106–2111, 2009. 6

[28] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014. 2

[29] Y. Wu, F. Yang, and H. Ling. Privacy-Protective-GAN for Face De-identification. *arXiv e-prints*, page arXiv:1806.08906, Jun 2018. 2

[30] J. J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial networks. 2017. 3

[31] X. Zhu, X. Liu, Z. Lei, and S. Z. Li. Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2