
LEGAN: Disentangled Manipulation of Directional Lighting and Facial Expressions by Leveraging Human Perceptual Judgements

Sandipan Banerjee

Affectiva, USA

sandipan.banerjee@affectiva.com

Ajjen Joshi

Affectiva, USA

ajjen.joshi@affectiva.com

Prashant Mahajan

New York University, USA

prashantmahajan70@gmail.com

Sneha Bhattacharya

Affectiva, USA

sneha.bhattacharya@affectiva.com

Survi Kyal

Affectiva, USA

survi.kyal@affectiva.com

Taniya Mishra

Affectiva, USA

taniya.mishra@affectiva.com

Abstract

Building facial analysis systems that generalize to extreme variations in lighting and facial expressions is a challenging problem that can potentially be alleviated using natural-looking synthetic data. Towards that, we propose LEGAN, a novel synthesis framework that leverages perceptual quality judgments for jointly manipulating lighting and expressions in face images, without requiring paired training data. LEGAN disentangles the lighting and expression subspaces and performs transformations in the feature space before upscaling to the desired output image. The fidelity of the synthetic image is further refined by integrating a perceptual quality estimation model into the LEGAN framework as an auxiliary discriminator. The quality estimation model is learned from face images rendered using multiple synthesis methods and their crowd-sourced naturalness ratings using a margin-based regression loss. Using objective metrics like FID and LPIPS, LEGAN is shown to generate higher quality face images when compared with popular GAN models like pix2pix, CycleGAN and StarGAN for lighting and expression synthesis. We also conduct a perceptual study using images synthesized by LEGAN and other GAN models, trained with and without the quality based auxiliary discriminator, and show the correlation between our quality estimation and visual fidelity. Finally, we demonstrate the effectiveness of LEGAN as training data augmente for expression recognition and face verification tasks.

1 Introduction

Deep learning [53] has engendered tremendous progress in automated facial analysis, with applications ranging from face verification [68, 21, 10] to expression classification [90]. However, building robust and accurate models that generalize effectively in-the-wild is still an open problem. A major part of this problem stems from training datasets failing to represent the “true” distribution of real world data [50, 64, 91, 5] (e.g. extreme lighting conditions [15, 69, 35]); or the training set may be non-uniformly distributed across the different classes, leading to the long-tail problem [58]. This skews the trained model’s representations towards the abundant classes and causes poor performance on test samples from the sparse classes [58].

One way to mitigate the imbalance problem, shown to work in multiple domains [56, 8, 13, 98], is to introduce synthetic samples into the training set. Many approaches for generating synthetic data exist [16, 7, 39], none as successful as GANs [34] in generating realistic face images [14, 47, 48, 24].



Figure 1: LEGAN jointly manipulates lighting and expression in face images while preserving subject identity.

Thus, in this work, we design a GAN model for synthesizing new views of an existing face image with the desired illumination and facial expression, while keeping the subject identity and other attributes constant. These synthetic images, when used as supplemental training data, can help build facial analysis systems that better generalize across variations in illumination and expressions.

One drawback of GAN-based face generation is the absence of an accurate and automated metric to judge the perceptual quality of synthesized images [52, 18]. In order to solve this problem, we also introduce a quality estimation model that can serve as a cheap but efficient proxy for human judgment while evaluating naturalness of synthetic face images. Instead of generating a single score for a distribution of synthetic images [105, 42, 76] or for image pairs [103, 72], our goal is to infer an image-quality score on a continuum for a single synthetic face image. With this in mind, we run an Amazon Mechanical Turk (AMT) experiment where turkers are instructed to score the naturalness of synthetic face images, generated using different 3D-model [8] and GAN-based [47, 48, 75, 2] synthesis approaches. We then build a feed forward CNN to learn representations from these images that map to their corresponding perceptual rating, using a margin based regression loss.

In addition to a traditional discriminator [34], this trained quality model is then used as an auxiliary discriminator in the synthesis framework, named LEGAN (Lighting-Expression GAN), that we propose in this paper. Instead of intertwining the two tasks [24], LEGAN decomposes the lighting and expression sub-spaces using a pair of hourglass networks (encoder-decoder) that generate transformation masks capturing the spatial changes required for target generation. The desired output image is then synthesized by a third hourglass network from these two masks.

To demonstrate the effectiveness of LEGAN, we qualitatively and quantitatively compare its synthesized images with those produced by three popular GAN based models [44, 106, 48] using objective metrics like FID [42], LPIPS [103], SSIM [93] and face match score. We also conduct a human rater study to evaluate the perceptual quality of LEGAN’s images and the contribution of the quality based auxiliary discriminator towards hallucinating perceptually superior and sharper images. Finally, we show the efficacy of LEGAN, when used as training data augmenter, in improving the generalizability of face verification and expression recognition models.

2 Related Work

Face Synthesis: Early approaches [16, 62] focused on stitching together similar looking facial patches from a gallery to synthesize a new face. Manipulating the facial shape using 3D models [45, 87, 58, 8] or deep features [25, 89] is another popular approach to generate new views. In recent times however researchers have pre-dominantly focused on using GANs [34] for synthesis, where an upsampling generator hallucinates faces from a noise vector, either randomly sampled from a distribution [34, 74, 47, 48] or interpreted from a different domain [67]. An existing face can also be encoded and then upsampled to obtain the desired attributes [44, 43, 94, 9, 66, 28, 32, 106, 22, 24].

Editing Expressions: Research in this domain started with modeling skin-muscle movements [30] for different facial expressions or swapping facial patches based on visual proximity [16]. With the advent of 3D face models, researchers used static [58] or morphable models [17, 100] to manipulate facial expressions with a higher degree of realism. Recently, the use of VAEs [11] and adversarial image-to-image translation networks have become extremely popular for editing facial expressions [44, 106, 24], with or without paired data for training. Some of these models use attention masks [73], facial shape information [33, 31] or exemplar videos [83] to guide the model in this task.

Editing Lighting: While methods like histogram equalization [70, 107] and gamma correction can shift the global luminance distribution and color encoding of an image, they cannot manipulate the direction of the light source itself. An early method [92] utilizes spherical harmonic based morphable models to manipulate the directional lighting in 3D. In [23], local linear adjustments are performed on overlapping windows to change the lighting profile of an image. Deep learning based approaches have also been proposed where the reflectance, normal and lighting channels are disentangled and

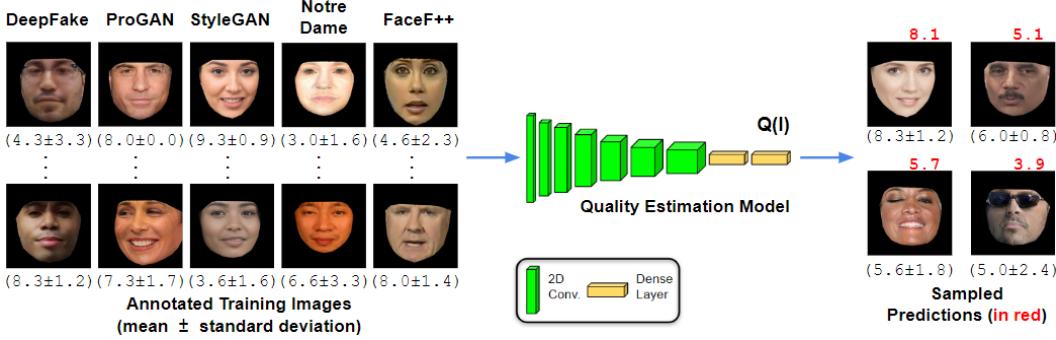


Figure 2: Our quality estimation model is trained on synthetic face images, varying in gender, ethnicity, lighting and facial attributes, generated with five methods [1, 47, 48, 8, 75] and their crowd-sourced perceptual ratings. To account for the subjective nature of human perception, our model utilizes a margin based regression loss to learn representations. This trained model can then generate naturalness predictions ($Q(I)$) of unseen face images.

edited to relight images [79, 82, 81, 38]. Alternatively, the desired lighting can be passed to an encoder-decoder pair as the target for lighting manipulation in the input image [104, 84, 60, 65].

Quality Estimation of Synthetic Face Images: Synthetic image quality is commonly evaluated using metrics like the Inception Score [76] or FID [42], which compare statistics of real and synthetic feature distributions, and output a single score for the whole distribution rather than the individual image. The features themselves are extracted from the Inception-v3 model [85], usually pre-trained on objects from [27], and not specifically faces. As these metrics do not take into account human judgements, they do not correlate well with perceptual realism [12, 18]. Consequently, researchers run perceptual studies to score the naturalness of synthetic images [102, 7]. These ratings are also used to design models that measure distortion between real and synthetic pairs [103, 72] or the coarse realism ('real' vs 'fake') of a synthetic image [105]. None of these evaluation models however are designed specifically for face images. Recently, [52] proposed a metric to rate the perceptual quality of a single image by using binary ratings from [105] as ground truth for synthetic face images generated by [47, 48]. Although this metric shows promise, their regression based model is trained on only 4,270 images and thus insufficient to reliably model the subjective nature of human judgements.

Unlike these methods, we build a synthetic face quality estimation model by leveraging perceptual ratings of over 37000 images generated using five different synthesis techniques [47, 48, 75, 1, 8]. Our quality model takes into account the variability in human judgements and generate a realism score for individual images rather than the whole set. We leverage this model as an auxiliary discriminator in the LEGAN framework for simultaneous lighting and facial expression manipulation. This novelty together with LEGAN's feature disentanglement improves the naturalness of the hallucinated images.

3 Quality Estimation Model

Our quality estimator model is trained with synthetic face images assembled and annotated in two sequential stages, as described below.

Stage I: We first generate 16,507 synthetic face images using the StyleGAN [48] generator. These images are then annotated by labelers using Amazon Mechanical Turk (AMT) on a scale of 0 - 10 for naturalness, where a 0 rating represents an unnatural image and 10 a hyper-realistic one. The images are then binned into two broad groups - 'unnatural' for AMT ratings between 0 - 5 and 'natural' for 5 - 10. We extract descriptors for each image from the 'avg_pool' layer of the ResNet50 [41] model, pre-trained on VGGFace2 [21] and train a linear SVM [26] with the extracted features of around 12,000 face images from this dataset and use the remaining 4,000 for parameter tuning. Post training, we use this SVM as a rough estimator of naturalness.

Stage II: In this stage, we perform the same AMT experiment again with a larger set of synthetic face images, collected from the following datasets:

1. **FaceForensics++**[75] - we randomly sample 1000 frames from this dataset consisting of 1000 video sequences that have been manipulated with four automated face manipulation methods.
2. **DeepFake**[2, 3] - we use sampled frames from 620 manipulated videos of 43 actors from [1].
3. **ProGAN** [47] - we generate 10,000 synthetic face images of non-existent subjects by training NVIDIA's progressively growing GAN model on the CelebA-HQ dataset [47].

Table 1: Detailed architecture of our quality estimation model Q (input size is $128 \times 128 \times 3$).

Layer	Filter/Stride/Dilation	# of filters
input	128×128	3
conv0	$4 \times 4 / 2 / 1$	64
conv1	$4 \times 4 / 2 / 1$	128
conv2	$4 \times 4 / 2 / 1$	256
conv3	$4 \times 4 / 2 / 1$	512
conv4	$4 \times 4 / 2 / 1$	1024
conv5	$4 \times 4 / 2 / 1$	2048
fc0	-	256
fc1	-	1

4. **StyleGAN** [48] - we extract 100,000 hyper-realistic face images of non-existent subjects generated using the StyleGAN model that were pre-filtered for quality [4].

5. **Notre Dame Synthetic Face Dataset** [8] - we randomly sample 163,000 face images, from the available 2M images, of synthetic subjects generated using ‘best-fitting’ 3D models.

To focus on near-frontal faces, we remove images with yaw over 15° in either direction, estimated using [40]. Since gender information is absent in most of the above datasets, we group the synthetic images using gender predictions from a pre-trained model [54]. Our trained SVM (from Stage I) is also used to rate the coarse naturalness of the collected images, using their ResNet50 features. We ensure balance in our synthetic dataset by sampling evenly from the natural and unnatural sets, as estimated by the SVM, and the perceived gender classes. To focus solely on the facial region, the pixels outside the convex hull formed by the facial landmarks, estimated using [19], of an image are masked. After the gender, facial yaw and naturalness based filtering, and the pre-processing step, we end up with 37,267 synthetic face images for our second AMT experiment.

Again, we ask Turkers to rate each image for naturalness on a scale of 0 - 10. Each image is shown to a Turker for 60 seconds to allow them time to make proper judgement even with slow network connection. We divide the full set of images into 72 batches such that each batch gets separately rated by 3 different Turkers. Post crowd-sourcing, we compute the mean (μ) and standard deviation (σ) from the 3 scores and assign them as the naturalness label for that image. We observe that the mean rating for the StyleGAN [48] images to be higher than that of the other methods. This is expected as the StyleGAN images we used are pre-filtered for realism. Some examples can be seen in Figure 2.

To train the quality estimation model, we use 80% of this annotated data and the rest for validation and testing. For augmentation, we only mirror the images as other techniques like translation, rotation and scaling drastically change their appearance compared to what the Turkers examined. Our model downsamples an input image using a set of strided convolution layers with Leaky ReLU [97] activation followed by two fully connected layers with linear activation and outputs a single realness scoring. Since both μ and σ are passed as image labels, we try to capture the inconsistency in the AMT ratings (i.e. the subjective nature of human perception) by formulating a margin based loss for training. The model weights are tuned such that its prediction is within an acceptable margin, set to σ , from the mean rating μ of the image. The loss L_N can be represented as:

$$L_N = \frac{1}{n} \sum_{i=1}^n \left\| \sigma - \|\mu - Q(I_i)\|_2^2 \right\|_2^2 \quad (1)$$

where n is the batch size, $Q(I_i)$ is the model prediction for the i -th image in the batch. Since the model is trained on the mean rating μ (regression) as the target rather than fixed classes (classification), L_N pushes the model predictions towards the confidence margin σ from μ .

Architecture Details: We share details of the architecture of our quality estimator Q in Table 1. The fully connected layers in Q are denoted as ‘fc’ while each convolution layer, represented as ‘conv’, is followed by Leaky ReLU [97] activation with a slope of 0.01.

Naturalness Rating Distribution in Training: In this section, we share the distribution of the naturalness ratings that we collected from the Amazon Mechanical Turk (AMT) experiment (Stage II). To do this, we average the perceptual rating for each synthetic face image from its three scores and increment the count of a particular bin in $[(0 - 1), (1 - 2), \dots, (8 - 9), (9 - 10)]$ based on the mean score. As described earlier, we design the AMT task such that a mean rating between 0 and 5 suggests the synthetic image to look ‘unnatural’ while a score between 5 and 10 advocates for its

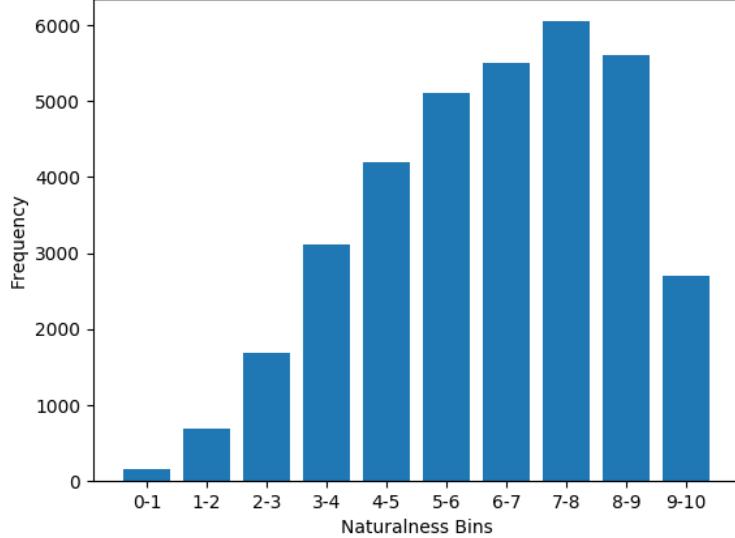


Figure 3: Histogram depicting the number of images in each naturalness bin, as rated by the Amazon Mechanical Turkers. Much more images fell on the ‘natural’ half (5 - 10) rather than the ‘unnatural’ one (0 - 5), suggesting the synthetic face images used in our study to be more or less realistic.

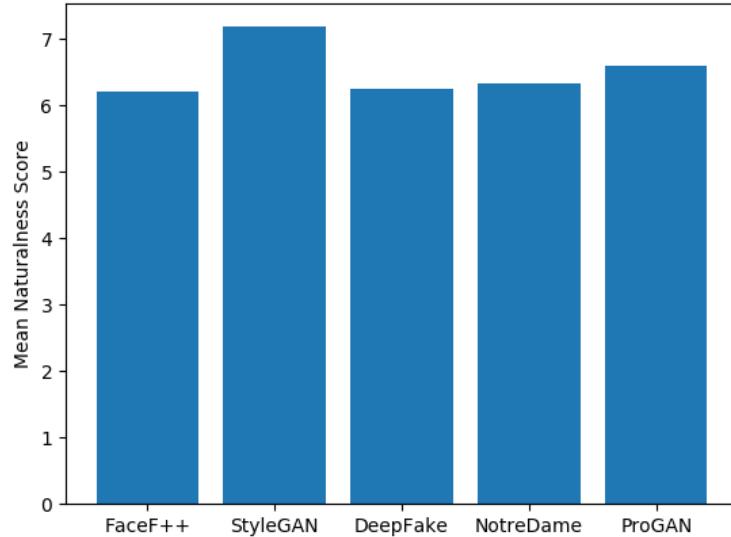


Figure 4: Mean naturalness rating of the different synthesis approaches used in our study [47, 48, 2, 75, 8]. As expected, the StyleGAN [48] images are rated higher than others as they were pre-filtered for quality[4].

naturalness. As can be seen in Figure 3, majority of the synthetic images used in our study generates a mean score that falls on the ‘natural’ side, validating their realism. When used to train our quality estimation model Q , these images tune its weights to look for the same perceptual features in other images while rating their naturalness.

To further check the overall perceptual quality of each of the different synthesis approaches used in our study [47, 48, 2, 75, 8], we separately find the mean rating for each synthetic face image generated by that method, depicted in Figure 4. It comes as no surprise for the StyleGAN [48] images to rank the highest, with a mean score over 7, as its face images were pre-filtered for quality [4]. The other four approaches perform roughly the same, generating a mean score that falls between 6 and 7.

Prediction Accuracy During Testing: As discussed before, we hold out 10% of the crowd-sourced data (3,727 face images) for testing our quality estimation model Q post training. Since Q never encountered these images during training, we use them to evaluate the effectiveness of our model. We separately compute the mean naturalness score for each synthesis approach used in our study

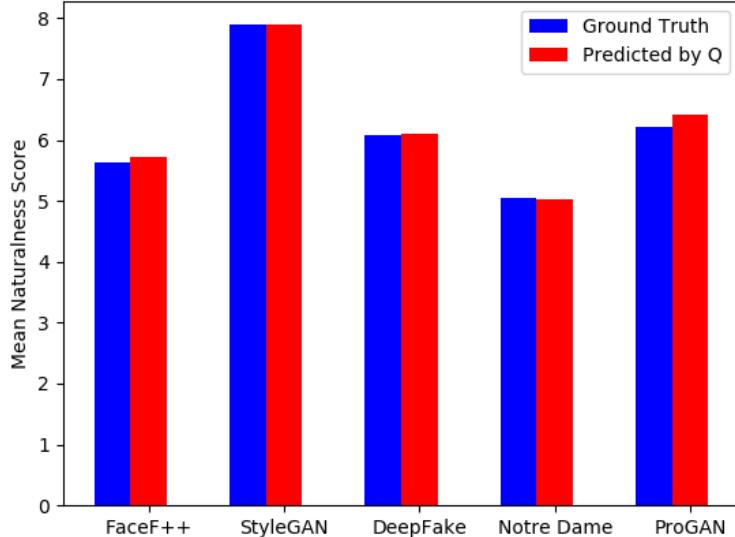


Figure 5: Mean naturalness rating, as estimated by Turkers (blue) and predicted by our trained quality estimation model Q (red), for the different synthesis approaches used in our study [47, 48, 2, 75, 8]. These ratings are specifically for images from the test split in our experiments, so Q never encountered them during training. Yet, Q is able to predict the naturalness of these images with a high degree of certainty.

and compare this value with the average quality score as predicted by Q . The results can be seen in Figure 5. Overall, our model predicts the naturalness score for each synthesis method with a high degree of certainty. Some qualitative results can also be seen in Figure 6.



Figure 6: Perceptual quality predictions by our trained quality estimation model (Q) on sample test images generated using [48, 47, 2, 75, 8]. For each image, the (mean \pm standard deviation) of the three naturalness scores, collected from AMT, is shown below while Q 's prediction is shown above in red.

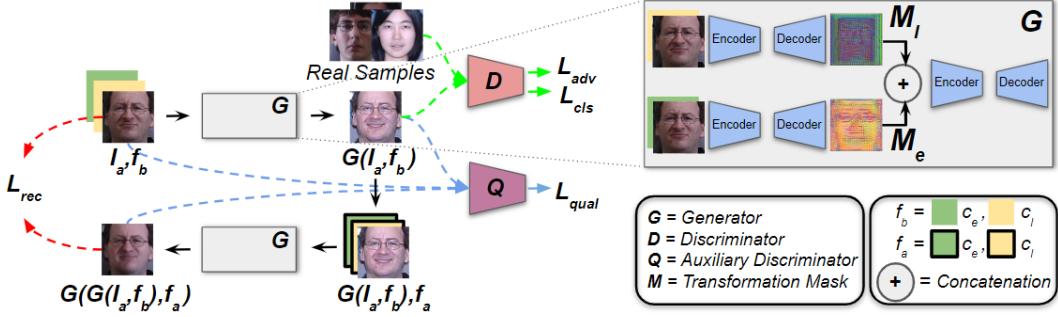


Figure 7: During training, G takes a face image I_a and target expression and lighting features f_b as input and disentangles the feature sub-spaces using a pair of hourglass networks (encoder-decoder) generating transformation masks M_e and M_l that are concatenated and passed through a third hourglass to hallucinate the target output $G(I_a, f_b)$. To eliminate the need of paired training data, we augment the generated output $G(I_a, f_b)$ with the source expression and lighting features f_a and pass it through the same generator to reconstruct the input $G(G(I_a, f_b), f_a)$, from which we compute the reconstruction error L_{rec} . Moreover, With the help of D and the quality based auxiliary discriminator Q , we calculate the adversarial (L_{adv}), feature classification (L_{cls}) and quality (L_{qual}) losses respectively. During testing, only G is required to generate synthetic images.

4 LEGAN

We design LEGAN to work with unpaired data, where the different lighting conditions and facial expressions need not be present for the same subject, and perform many-to-many domain transformations i.e. any input lighting-expression to any output lighting-expression. We describe the architecture and objective functions of LEGAN in this section, an overview of which can be seen in Figure 7.

4.1 Architecture

Generator: Our generator G , composed of three hourglass networks (encoder-decoder), starts with an input RGB face image I_a and a target attributes vector f_b that corresponds to expression and lighting conditions c_e and c_l respectively. The first hourglass receives I_a concatenated with c_e while the second one receives I_a concatenated with c_l , thus disentangling the transformation task. Inside each hourglass, the concatenated tensor is downsampled using strided convolutions and then passed through a set of residual blocks [41] before being upsampled using pixel shuffling layers [80], a choice that we justify in the supplementary text. Each convolution layer is followed by instance normalization [88] and ReLU activation [59] for non-linearity. These upsampled images are the transformation masks M_e and M_l that map the changes in pixel intensity required to translate I_a to conditions specified in f_b . M_e and M_l are concatenated together and fed to the third hourglass to generate the output image $G(I_a, f_b)$. The objective of dividing the generation process into two stages and hallucinating the transformation masks is two fold - (a) easing the task of each hourglass by simply making it focus on registering the required expression or lighting changes instead of both registration and hallucination, and (b) making the transformation process more explainable, with salient pixels prominent in M_e and M_l , as can be seen in Figure 8.

Since G is composed of three hourglass networks, we separately describe their architecture in Tables 2, 3 and 4 respectively. The convolution layers, residual blocks and pixel shuffling layers are indicated as ‘conv’, ‘RB’, and ‘PS’ respectively in the tables. After each of ‘conv’ and ‘PS’ layer in an hourglass, we use ReLU activation and instance normalization [88], except for the last ‘conv’ layer where a tanh activation is used [74, 77].

Discriminator: The discriminator D takes the output image $G(I_a, f_b)$ and predicts not only its realness score but also classifies its attributes f_b . D is composed of strided convolution layers with Leaky ReLU [97] activation that downsample the image to extract its encoded feature map. We use a patch discriminator [44] that takes this encoded feature map and passes it through a single channel convolution to get the realness map D_{src} . This feature map is also operated by a convolution layer with k filters to get the attributes prediction map D_{cls} , where $k = \text{number of channels in } f_b$.

The detailed description of D can be found in Table 5. Similar to Q , each convolution layer is followed by Leaky ReLU [97] activation with a slope of 0.01 in D , except for the final two convolution layers that output the realness matrix D_{src} and the classification map D_{cls} .

Table 2: Hourglass architecture for expression mask (M_e) synthesis in the generator G . The input size is $128 \times 128 \times 9$, three RGB channels (I_a) and six expression channels (c_e).

Layer	Filter/Stride/Dilation	# of filters
input	$128 \times 128 / -/-$	9
conv0	$7 \times 7 / 1 / 1$	64
conv1	$4 \times 4 / 2 / 1$	128
conv2	$4 \times 4 / 2 / 1$	256
RB0	$3 \times 3 / 1 / 1$	256
RB1	$3 \times 3 / 1 / 1$	256
RB2	$3 \times 3 / 1 / 1$	256
RB3	$3 \times 3 / 1 / 1$	256
RB4	$3 \times 3 / 1 / 1$	256
RB5	$3 \times 3 / 1 / 1$	256
PS0	-	256
conv3	$4 \times 4 / 1 / 1$	128
PS1	-	128
conv4	$4 \times 4 / 1 / 1$	64
conv5 (M_e)	$7 \times 7 / 1 / 1$	3

Table 3: Hourglass architecture for lighting mask (M_l) synthesis in the generator G . The input size is $128 \times 128 \times 23$, three RGB channels (I_a) and twenty expression channels (c_l).

Layer	Filter/Stride/Dilation	# of filters
input	$128 \times 128 / -/-$	23
conv0	$7 \times 7 / 1 / 1$	64
conv1	$4 \times 4 / 2 / 1$	128
conv2	$4 \times 4 / 2 / 1$	256
RB0	$3 \times 3 / 1 / 1$	256
RB1	$3 \times 3 / 1 / 1$	256
RB2	$3 \times 3 / 1 / 1$	256
RB3	$3 \times 3 / 1 / 1$	256
RB4	$3 \times 3 / 1 / 1$	256
RB5	$3 \times 3 / 1 / 1$	256
PS0	-	256
conv3	$4 \times 4 / 1 / 1$	128
PS1	-	128
conv4	$4 \times 4 / 1 / 1$	64
conv5 (M_l)	$7 \times 7 / 1 / 1$	3

Table 4: Hourglass architecture for target image ($G(I_a, f_b)$) synthesis in the generator G . The input size is $128 \times 128 \times 6$, three expression mask channels (M_e) and three lighting mask channels (M_l).

Layer	Filter/Stride/Dilation	# of filters
input	$128 \times 128 / -/-$	6
conv0	$7 \times 7 / 1 / 1$	64
conv1	$4 \times 4 / 2 / 1$	128
conv2	$4 \times 4 / 2 / 1$	256
RB0	$3 \times 3 / 1 / 1$	256
RB1	$3 \times 3 / 1 / 1$	256
RB2	$3 \times 3 / 1 / 1$	256
RB3	$3 \times 3 / 1 / 1$	256
RB4	$3 \times 3 / 1 / 1$	256
RB5	$3 \times 3 / 1 / 1$	256
PS0	-	256
conv3	$4 \times 4 / 1 / 1$	128
PS1	-	128
conv4	$4 \times 4 / 1 / 1$	64
conv5 ($G(I_a, f_b)$)	$7 \times 7 / 1 / 1$	3

Table 5: Detailed architecture of LEGAN’s discriminator D (input size is $128 \times 128 \times 3$).

Layer	Filter/Stride/Dilation	# of filters
input	128×128	3
conv0	$4 \times 4 / 2 / 1$	64
conv1	$4 \times 4 / 2 / 1$	128
conv2	$4 \times 4 / 2 / 1$	256
conv3	$4 \times 4 / 2 / 1$	512
conv4	$4 \times 4 / 2 / 1$	1024
conv5	$4 \times 4 / 2 / 1$	2048
conv6 (D_{src})	$3 \times 3 / 1 / 1$	1
conv7 (D_{cls})	$1 \times 1 / 1 / 1$	26



Figure 8: Sample LEGAN results ($G(I_a, f_b)$), with generated expression (M_e) and lighting (M_l) masks for input face images (I_a) with different color composition and levels of shadow. The salient pixels for synthesizing the target expression and lighting automatically ‘heat up’ in M_e and M_l respectively, similar to flow maps.

Auxiliary Discriminator: We integrate the perceptual quality model Q , described in Section 3, into the LEGAN model graph to further refine the naturalness of the images synthesized by G . Unlike D , we do not train Q jointly with G but use the weights of a pre-trained snapshot.

4.2 Loss Function

1. Adversarial Loss: D is trained to distinguish a real face image I_a from its synthetic counterpart and judge the realness of the hallucinated image $G(I_a, f_b)$. To stabilize the gradients and improve quality, we use the WGAN [6] based objective L_{adv} for this task with a gradient penalty [37], set as:

$$L_{adv} = \mathbb{E}_{I_a}[D_{src}(I_a)] - \mathbb{E}_{I_a, f_b}[D_{src}(G(I_a, f_b))] - \lambda_{gp}\mathbb{E}_{\hat{I}}[(\|\nabla_{\hat{I}} D_{src}(\hat{I})\|_2 - 1)^2] \quad (2)$$

where \hat{I} is sampled uniformly from real and synthetic images and λ_{gp} is a tunable parameter. While D tries to minimize this to separate the synthetic from the real, G tries to maximize it by fooling D .

2. Classification Loss: To ensure the target lighting and expression are correctly rendered by G and enable LEGAN to do many-to-many translations, we formulate a classification loss using D ’s predictions, in the form of D_{cls} . The loss L_{cls} is computed as:

$$L_{cls} = \mathbb{E}_{I_a, f_a}[-\log D_{cls}(f_a | I_a)] + \mathbb{E}_{I_a, f_b}[-\log D_{cls}(f_b | G(I_a, f_b))] \quad (3)$$

where f_a and f_b are the original and target attributes of an input image I_a .

3. Reconstruction Loss: To preserve the subject identity without using paired data, we use a cyclic reconstruction loss [106] L_{rec} between I_a and its reconstruction $G(G(I_a, f_b), f_a)$, computed as:

$$L_{rec} = \mathbb{E}_{I_a, f_b, f_a}[\|I_a - G(G(I_a, f_b), f_a)\|_1] \quad (4)$$

4. Quality Loss: We use Q ’s predictions for further improving with perceptual realism of the synthetic images. Masked versions of the input image I_a' , the synthesized output $G(I_a, f_b)'$ and the reconstructed input $G(G(I_a, f_b), f_a)'$, produced using facial landmarks extracted by [19], are used for loss computation as follows:

$$L_{qual} = \mathbb{E}_{I_a, f_b}[\|q - Q(G(I_a, f_b))\|_1] + \mathbb{E}_{I_a, f_b, f_a}[\|Q(I_a') - Q(G(G(I_a, f_b), f_a))\|_1] \quad (5)$$

where q is a hyper-parameter that can be tuned to lie between 5 (realistic) and 10 (hyper-realistic). We set $q = 8$, a choice that we justify in the next section.

Full Loss: We also apply total variation loss [46] L_{tv} on $G(I_a, f_b)$ and $G(G(I_a, f_b), f_a)$ to smooth boundary pixels and set the final training objective L as a weighted sum of the five losses as:

$$L = L_{adv} + \lambda_{cls}L_{cls} + \lambda_{rec}L_{rec} + \lambda_{qual}L_{qual} + \lambda_{tv}L_{tv} \quad (6)$$

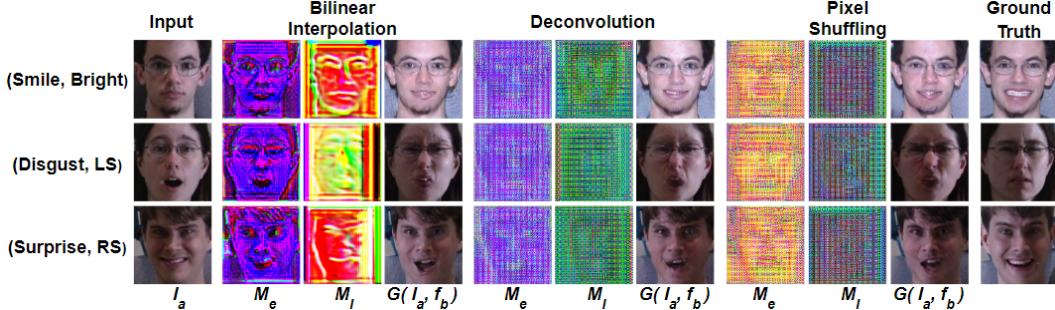


Figure 9: Adding different upsampling techniques in our decoder modules generates hallucinations $G(I_a, f_b)$ with slightly different perceptual scores for the same input I_a . However, the transformation masks M_e and M_l are smoother and more meaningful when bilinear interpolation is used for upsampling. Since both deconvolution [101] and pixel shuffling [80] learn the intensity of the upsampled pixels instead of simple interpolation, the masks they generate are more fragmented and discrete. We use pixel shuffling in our final LEGAN model.

Table 6: Effects of different upsampling - quantitative results on held out CMU-MultiPIE [36] test set.

Metrics	Bilinear Interpolation	Deconvolution [101]	Pixel Shuffling [80]
FID [42] \downarrow	37.75	41.27	38.40
LPIPS [103] \downarrow	0.140	0.132	0.131
SSIM [93] \uparrow	0.565	0.578	0.584
Match Score [41, 21] \uparrow	0.608	0.626	0.617

5 Experiments and Results

Training Data: We utilize 36,657 frontal RGB images from the CMU-MultiPIE dataset [36], with 20 different lighting conditions and 6 acted facial expressions, to build our model. For training we use 33,305 images of 303 subjects and the remaining 3,352 images of 34 subjects for testing. The training data is highly skewed towards ‘Neutral’ and ‘Smile’ compared to the other 4 expressions but the distribution is almost uniform for the lighting classes. We align each image using their eye landmarks extracted with [19] and resize to $128 \times 128 \times 3$. We do not fine-tune LEGAN on any other data and solely rely on its generalizability for the different experimental tasks.

Implementation Details: To learn the model, we use the Adam optimizer [51] with a learning rate of 0.0001 and parameters β_1 and β_2 set to 0.5 and 0.999 respectively. The different loss weights λ_{gp} , λ_{cls} , λ_{rec} , λ_{qual} and λ_{tv} are set empirically to 10, 20, 10, 0.5 and 0.0000001 respectively. As done in [73], we train D 5 times for each training iteration of the G . The model is trained with a batch size of 4 on a single NVIDIA Titan Xp GPU for 50 epochs, which takes around 28 hours to finish.

Optimal Upsampling: To check the effect of the different upsampling approaches on hallucination quality, we separately apply bilinear interpolation, deconvolution [101] and pixel shuffling [80] on the decoder module of the three hourglass networks in LEGAN’s generator G . While the upsampled pixels are interpolated based on the original pixel in the first approach, the other two approaches explicitly learn the possible intensity during upsampling. More specifically, pixel shuffling blocks learn the intensity for the pixels in the fractional indices of the original image (i.e. the upsampled indices) by using a set convolution channels and have been shown to generate sharper results than deconvolution. Unsurprisingly, it generates the best quantitative results by outperforming the other two upsampling approaches on 2 out of our 4 objective metrics, as shown in Table 6. Hence we use pixel shuffling blocks in our final implementation of LEGAN.

However, as can be seen in Figure 9, the expression and lighting transformation masks M_e and M_l are more meaningful when interpolated rather than explicitly learned. This interpolation leads to a smoother flow of upsampled pixels with facial features and their transformations visibly more noticeable compared to deconvolution and pixel shuffling.

Optimal Value of q : As discussed before, we set the value of the hyper-parameter $q = 8$ for computing the quality loss L_{qual} . We arrive at this specific value after experimenting with different possible values. Since q acts as a target for perceptual quality while estimating L_{qual} during the forward pass, it can typically range from 5 (realistic) to 10 (hyper-realistic). We set q to all possible integral values between 5 and 10 for evaluating the synthesis results both qualitatively (Figure 10) and quantitatively (Table 7).

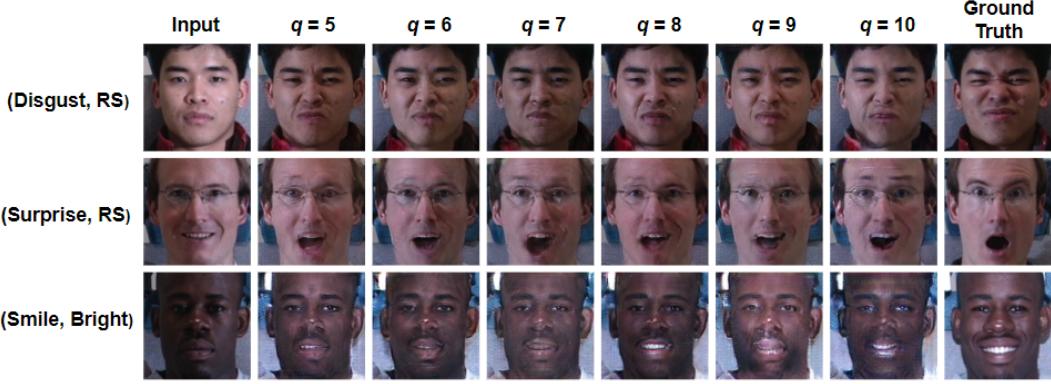


Figure 10: Sample results illustrating the effect of the hyper-parameter q on synthesis quality. Since it generates more stable and noticeable expressions (bottom row) with fewer artifacts (middle row), we set $q = 8$ for the final LEGAN model.

Table 7: Optimal value of q - quantitative results on held out CMU-MultiPIE [36] test set.

Metrics	$q = 5$	$q = 6$	$q = 7$	$q = 8$ (LEGAN)	$q = 9$	$q = 10$
FID [42] \downarrow	46.01	42.58	44.29	38.40	53.45	47.91
LPIPS [103] \downarrow	0.139	0.138	0.140	0.131	0.146	0.148
SSIM [93] \uparrow	0.559	0.560	0.572	0.584	0.565	0.545
Match Score [41, 21] \uparrow	0.599	0.599	0.601	0.617	0.596	0.557

As can be seen, when q is set to 8, LEGAN generates more stable images with much less artifacts compared to other values of q . Also, the synthesized expressions are visibly more noticeable for this value of q (Figure 10, bottom row). When evaluated quantitatively, images generated by LEGAN with $q = 8$ garner the best score for all 4 objective metrics. This is interesting as setting $q = 10$ (and not 8) should ideally generate hyper-realistic images and consequently produce the best quantitative scores. We attribute this behavior of LEGAN to the naturalness distribution of the images used to train our quality model Q . Since majority of these images fell in the (7-8) and (8-9) bins, and very few in (9-10) (as shown in Figure 3), Q 's representations are aligned to this target. As a result, Q tends to rate hyper-realistic face images (i.e. images with mean naturalness rating between 8 - 10) with a score around 8. Such an example can be seen in the rightmost column of the first row in Figure 6, where Q rates a hyper-realistic StyleGAN generated image [48] as 8.3. Thus, setting $q = 8$ for L_{qual} computation (using trained Q 's weights) during LEGAN training produces the optimal results.

Comparison with Other GAN Models: To compare with LEGAN, we choose 3 popular GAN models - pix2pix [44], CycleGAN [106], and StarGAN [24]. We train these models for lighting and expression manipulation with the same MultiPIE [36] training split for 50 epochs. Unlike LEGAN and StarGAN that can perform many-to-many translations, pix2pix (requires paired data) and CycleGAN are only capable of doing one-to-one mappings. Hence, we train 9 different versions of pix2pix and CycleGAN for 9 specific lighting-expression translations:

1. ('Neutral', 'Left Shadow') \rightarrow ('Smile', 'Bright')
2. ('Neutral', 'Left Shadow') \rightarrow ('Disgust', 'Right Shadow')
3. ('Neutral', 'Left Shadow') \rightarrow ('Smile', 'Dark')
4. ('Neutral', 'Left Shadow') \rightarrow ('Smile', 'Right Shadow')
5. ('Smile', 'Left Shadow') \rightarrow ('Surprise', 'Right Shadow')
6. ('Surprise', 'Left Shadow') \rightarrow ('Disgust', 'Left Shadow')
7. ('Surprise', 'Left Shadow') \rightarrow ('Disgust', 'Bright')
8. ('Surprise', 'Left Shadow') \rightarrow ('Smile', 'Left Shadow')
9. ('Surprise', 'Left Shadow') \rightarrow ('Smile', 'Right Shadow')

We pick difficult transformation tasks, like ('Surprise' \rightarrow 'Disgust'), intentionally to test LEGAN's capability of performing such translations without seeing any paired data during training. Although this puts these two models, especially pix2pix, at an advantage, we treat them as benchmarks to compare with the unpaired, many-to-many models LEGAN and StarGAN, which are capable of doing the full range of $(6 \times 20)^2$ expression-lighting translations possible in MultiPIE [36]. Additionally,

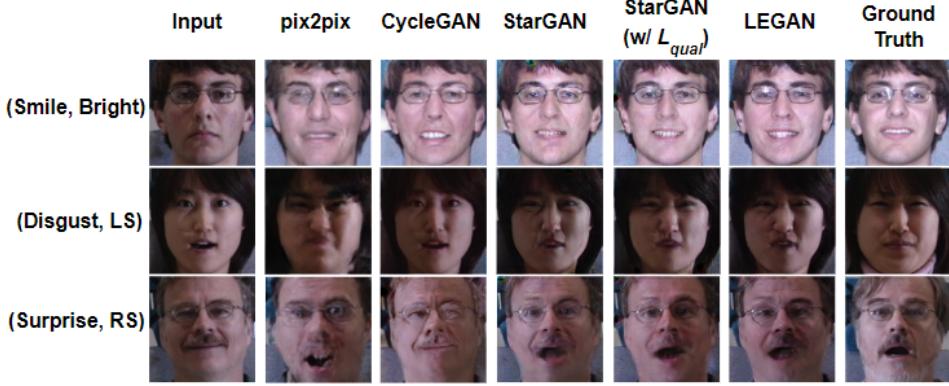


Figure 11: Sample results on MultiPIE [36] test images comparing LEGAN with popular GAN models. The target conditions are shown at the start of each row (LS = Left Shadow, RS = Right Shadow). LEGAN hallucinates subtle muscle movements like frowns more prominently (middle row) while preserving the subject identity.

Table 8: Quantitative comparison with popular GAN models on held out CMU-MultiPIE [36] test set.

Metrics	pix2pix [44]	CycleGAN [106]	StarGAN [24]	StarGAN w/ L_{qual}	LEGAN
FID [42]↓	43.64	41.71	50.42	41.73	38.40
LPIPS [103]↓	0.257	0.142	0.144	0.145	0.131
SSIM [93]↑	0.614	0.532	0.556	0.572	0.584
Match Score [41, 21]↑	0.506	0.592	0.603	0.608	0.617
Human Preference ↑	33.26%	9.94%	10.16%	22.24%	24.38%

to gauge the effect of L_{qual} on off-the-shelf models we train StarGAN separately with the auxiliary discriminator added.

After training, we compare the synthetic image generated by each model with the corresponding target image (ground truth) using the following objective metrics - (1) FID [42] and (2) LPIPS [103] to gauge the realism, (3) SSIM [93] to measure noise, and (4) face matching score using pre-trained ResNet50 [41, 21] features and Pearson correlation coefficient. We also run a perceptual study using face images generated by these models where we ask 35 non-expert human raters to pick an image from a lineup that best matches - (1) a target facial expression and (2) a target lighting condition, for MultiPIE [36] subjects. The raters are first shown real examples of the target expressions (Neutral, Smile, Surprise and Disgust) and lighting conditions (Bright, Dark, Left Shadow and Right Shadow). The raters are shown a total of 54 rows of images, where each row consists of an actual image of the subject with bright lighting and neutral expression and the same subject synthesized for the target expression and lighting by the 5 GAN models, presented in a randomized order. We aggregate the rater votes across all rows and normalize them for each model. The results are shown in Table 8.

As can be seen, LEGAN synthesizes perceptually superior face images (FID, LPIPS) while retaining subject identity (match score) better than pix2pix and the other GAN models, even without using paired data for training. Although visually not as sharp, we find pix2pix to generate target facial expressions with a higher intensity compared to LEGAN, which ranks second, as quantified by the SSIM and human preference scores. This comes as no surprise, as pix2pix explicitly learns translations from the paired base and target images during training. Since it only does one-to-one translations, pix2pix's model weights are also more tuned for that particular task compared to LEGAN. However, this setting requires training N versions of pix2pix for N different translation tasks making it impractical for real world applications. We also find adding L_{qual} to StarGAN improves almost all its metric scores underpinning the value our quality estimator Q even when coupled with off-the-shelf models. On top of enhancing the overall sharpness, L_{qual} removes bullet-hole artifacts, similar to [49], from the peripheral regions of the output face. Such an artifact can be seen in (row 1, col 4, forehead-hair boundary) of Figure 11, which is eliminated by adding L_{qual} (row 1, col 5).

Perceptual Study Details: We share more details about the interface used for our perceptual study here. As shown in Figure 12, we ask the raters to pick the image that best matches a target expression and lighting condition. To provide a basis for making judgement, we also share a real image of the same subject with neutral expression and bright lighting condition. However, this is not necessarily

26. For the person in the left-most image, which of the 5 images on the right best represents that person with the following 2 conditions: facial expression=DISGUST and illumination condition=RIGHT SHADOW?



Figure 12: Our perceptual study interface: given a base face image with neutral expression and bright lighting (leftmost image), a rater is asked to select the image that best matches the target expression ('Disgust') and lighting ('Right Shadow') for the same subject.

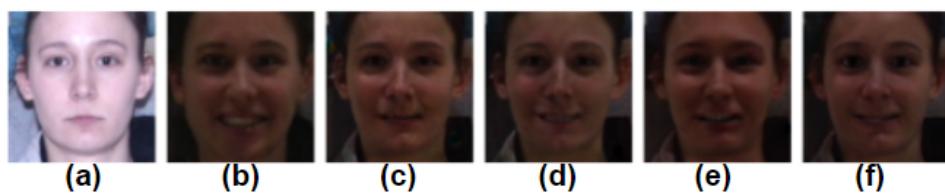


Figure 13: Instance where, given the base face image (a) majority of the raters picked the image generated by pix2pix (b) over CycleGAN (e), StarGAN (c), StarGAN w/ L_{qual} (d) and LEGAN (f). The target expression and lighting were 'Smile' and 'Dark' respectively.



Figure 14: Instance where, given the base face image (a) majority of the raters picked the image generated by CycleGAN (d) over pix2pix (b), StarGAN (e), StarGAN w/ L_{qual} (c) and LEGAN (f). The target expression and lighting were 'Disgust' and 'Right Shadow' respectively.



Figure 15: Instance where, given the base face image (a) majority of the raters picked the image generated by StarGAN (f) over pix2pix (d), CycleGAN (c), StarGAN w/ L_{qual} (e) and LEGAN (b). The target expression and lighting were 'Surprise' and 'Right Shadow' respectively.



Figure 16: Instance where, given the base face image (a) majority of the raters picked the image generated by StarGAN w/ L_{qual} (f) over pix2pix (c), CycleGAN (d), StarGAN (e) and LEGAN (a). The target expression and lighting were 'Smile' and 'Right Shadow' respectively.

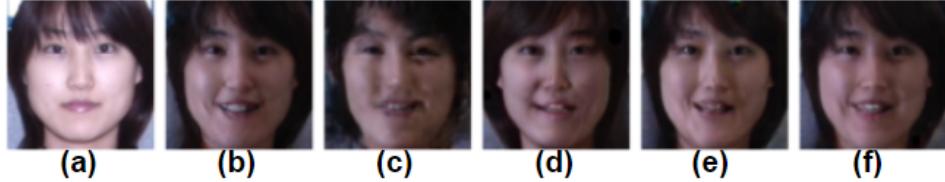


Figure 17: Instance where, given the base face image (a) majority of the raters picked the image generated by LEGAN (b) over pix2pix (c), CycleGAN (d), StarGAN (e) and StarGAN w/ L_{qual} (f). The target expression and lighting were ‘Smile’ and ‘Right Shadow’ respectively.

Table 9: Performance of LightCNN-29 model [96] on IJB-B [95] with and without LEGAN based augmentation.

Training Data	Real Images [99] (# Identities)	Synthetic Images (# Identities)	IJB-B [95] Performance (TPR@FPR = 0.01)
Original	439,999 (10,575)	0	0.952
Augmented	439,999 (10,575)	439,999 (10,575)	0.965

the input to the synthesis models for the target expression and lighting generation, as we want to estimate how these models do when the input image has more extreme expressions and lighting conditions. The image order is also randomized to eliminate any bias.

In Figures 13, 14, 15, 16 and 17 we show instances where the raters preferred images synthesized using pix2pix [44], CycleGAN [106], StarGAN [24], StarGAN w/ L_{qual} and LEGAN respectively. As can be seen, LEGAN generates sharp results with target expression and lighting while preserving the subject identity. Although pix2pix does not properly preserve the subject identity, it does generate more intense expressions compared to LEGAN and the other models. This can be attributed to the fact that pix2pix’s training requires paired data where the model sees real instance of both the input and the target output, unlike LEGAN. Hence pix2pix ranks highest overall, while LEGAN ranks second.

Table 10: Model [29] performance (ROC AUC) on AffectNet [61] with and without LEGAN based augmentation

Training Data	Real Images [61]	Synthetic Images	‘Neutral’	‘Happy’	‘Surprise’	‘Disgust’
Original	204,325	0	0.857	0.953	0.872	0.886
Augmented	204,325	279,324	0.858	0.956	0.886	0.895

Effectiveness as Training Data Augmenter: We examine the use of LEGAN as training data augmenter for face verification and expression recognition tasks using the IJB-B [95] and AffectNet [61] datasets. For face verification, we use the CASIA-WebFace dataset [99] and the LightCNN-29 [96] architecture due to their popularity in this domain. We randomly sample 439,999 images of 10,575 subjects from [99] for training and 54,415 images for validation. We augment the training set by randomly manipulating the lighting and expression of each image (Table 9, row 2). The LightCNN-29 model is trained from scratch separately with the original and augmented sets and its weights saved when validation loss plateaus across epochs. These saved snapshots are then used to extract features from a still image or video frame in the IJB-B [95] dataset. For each IJB-B template, a mean feature is computed using video and media pooling operations [56] and match score between such features is calculated with Pearson correlation. We find the model trained with the augmented data to improve upon the verification performance of the baseline (Table 9). This suggests that the LEGAN generated images retain their original subject identity and can boost the robustness of classification models towards intra-class variance in expressions and lighting .

For expression classification, we use a modified version of the AU-classification model from [29] (Leaky ReLU [97] and Dropout added) and manually annotated AffectNet [61] images for the (‘Neutral’, ‘Happy’, ‘Surprise’, ‘Disgust’) classes, as these 4 expressions overlap with MultiPIE [36]. The classification model is trained with 204,325 face images from AffectNet’s training split, which is highly skewed towards the ‘Happy’ class (59%) and has very few images for the ‘Surprise’ (6.2%) and ‘Disgust’ (1.6%) expressions. To balance the training distribution, we populate each sparse class with synthetic images generated by LEGAN from real images belonging to any of the other 3 classes. We use the original and augmented (balanced) data separately to train two versions of the model for expression classification. As there is no test split, we use the 2,000 validation images for testing, as done in other works [90]. We find the synthetic images, when used in training, to substantially

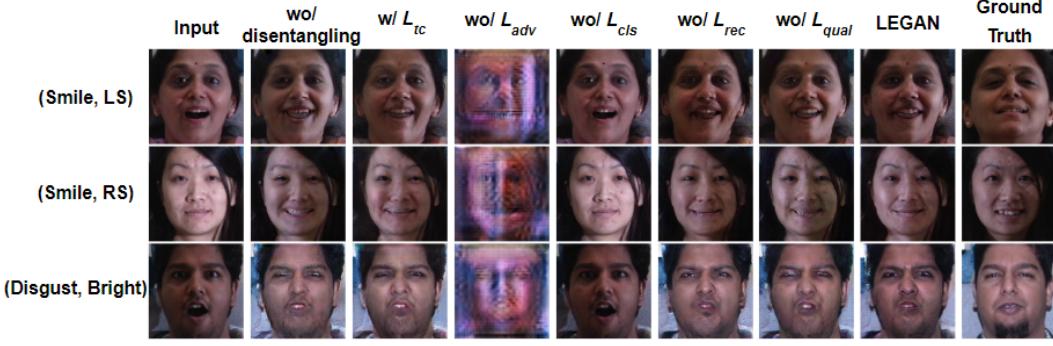


Figure 18: Sample qualitative results from LEGAN and its variants.

Table 11: Ablation studies - quantitative results on held out CMU-MultiPIE [36] test set.

Metrics	wo/ disentangling	w/ L_{tc} [78]	wo/ L_{adv}	wo/ L_{cls}	wo/ L_{rec}	wo/ L_{qual}	wo/ L_{qual}	LEGAN
FID [42] \downarrow	46.90	48.69	303.3	39.78	45.72	46.00	38.40	
LPIPS [103] \downarrow	0.144	0.140	0.528	0.221	0.138	0.141	0.131	
SSIM [93] \uparrow	0.575	0.558	0.285	0.321	0.565	0.578	0.584	
Match Score [41, 21] \uparrow	0.588	0.594	0.044	0.680	0.603	0.595	0.617	

improve test performance especially for the previously under-represented ‘Surprise’ and ‘Disgust’ classes (Table 10). This further validates the realism of the expressions generated by LEGAN.

Ablation Study: To analyze the contribution of each loss component on synthesis quality, we prepare 5 different versions of LEGAN by removing (feature disentanglement, L_{adv} , L_{cls} , L_{rec} , L_{qual}) from G while keeping everything else the same. To check the impact of the recently proposed triple consistency loss [78], we also prepare a version of LEGAN by adding L_{tc} to its objective. Instead of facial pose [78], we formulate L_{tc} as $\mathbb{E}_{I_a, f_b, f_c} [\|G(I_a, f_c) - G(G(I_a, f_b), f_c)\|_1]$, where f_c is a target expression and lighting vector different from f_b and f_a . The qualitative and quantitative results, produced using MultiPIE test data, are shown in Figure 18 and Table 11 respectively. As expected, we find L_{adv} to be crucial for realistic hallucinations while L_{rec} acts as a regularizer by suppressing artifacts and preserving subject identity. When L_{cls} is removed, LEGAN outputs the input image back as the target attributes are not checked by D anymore. Removing L_{qual} deteriorates the overall naturalness, with artifacts manifesting in the eye and mouth regions. Without the feature disentanglement, we find the synthetic images to be blurrier and contain unnatural facial features (especially the eyes and the teeth). Interestingly, we observe adding L_{tc} to dampen the facial expression when going from an intense expression to another (e.g. Surprise \rightarrow Disgust, Figure 18, bottom row). Since L_{tc} tries to minimize the distance between $G(I_a, f_c)$ and $G(G(I_a, f_b), f_c)$, it drives G ’s representations towards the space between f_a and f_b instead of moving them closer to f_b .

Model Limitations & Qualitative Results: Although LEGAN is trained on just frontal face images acquired in a controlled setting, it can still generate realistic new views even for non-frontal images with a variety of expressions, as shown in Figures 21 and 22. However, as with any synthesis model, LEGAN also has its limitations. In majority of the cases where LEGAN fails to synthesize a realistic image, the input expression is extreme with non-frontal head pose, as can be seen in Figure 19. In most of these input images, the original expression is rather extreme (tongue sticking out, mouth wide open, painting) and LEGAN fails to generalize to these particular images. One way to mitigate this is



Figure 19: Failure cases - (a) mouth not opened, (b) tongue still out (extreme input expression), (c) mouth not properly closed, (d) no change in expression (extreme input expression). For each pair, the input image is on the left and its synthetic version, generated by LEGAN for the specified target expression, is on the right.

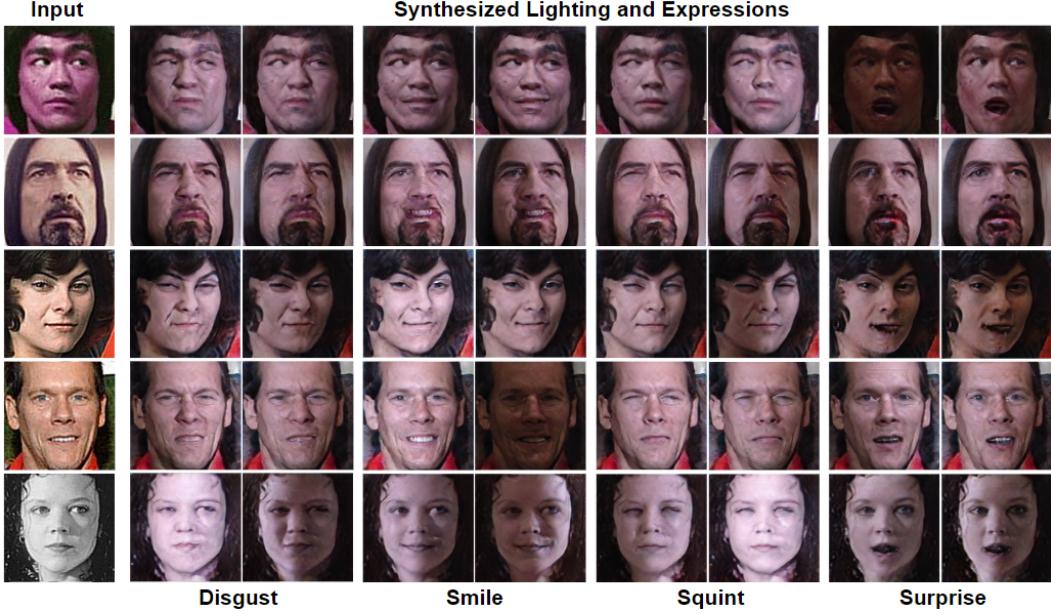


Figure 20: Different hallucination results for the same input image. The target expression can be seen under each result pair, second column onward. The two target lighting conditions however, exhibited in the two images in a pair, is randomly selected. All images are $128 \times 128 \times 3$.

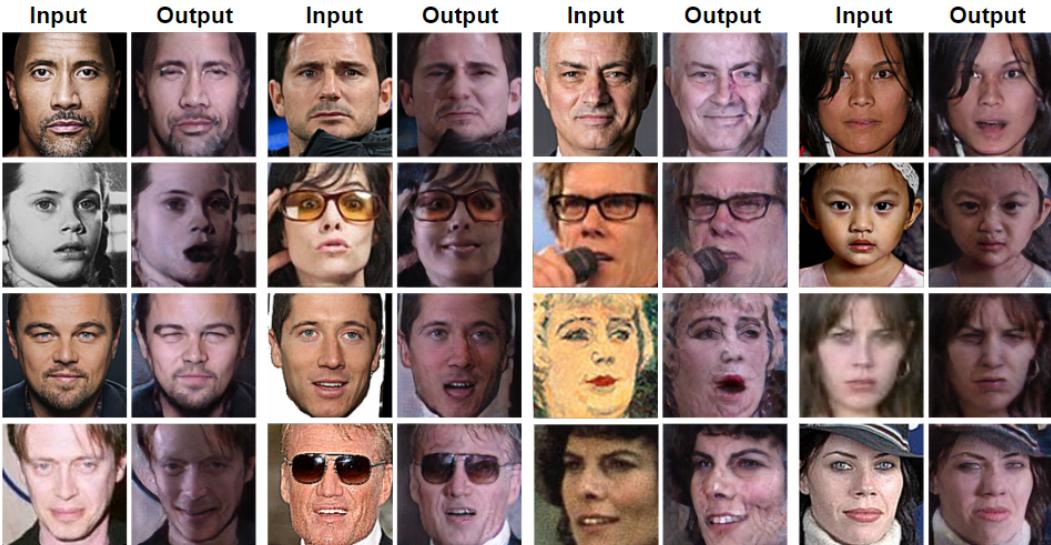


Figure 21: More qualitative results demonstrating the generalizability of LEGAN. For each input image, we generate the output image with a randomly selected target lighting and expression. The input images vary in gender, ethnicity, color composition, facial pose, expression, lighting and accessories. All images are $128 \times 128 \times 3$.

to formulate a hallucination pathway of individual facial action unit [86] manipulations required to translate one extreme expression to another.

We share more qualitative results generated by LEGAN on unconstrained data collected from the internet. In Figure 20, we show the different expression and lighting hallucination results for the same input image. Synthesis results, with a randomly selected target lighting and expression, on input face images that vary in gender, ethnicity, color composition, facial pose, expression, lighting and accessories can be seen in Figures 21 and 22.

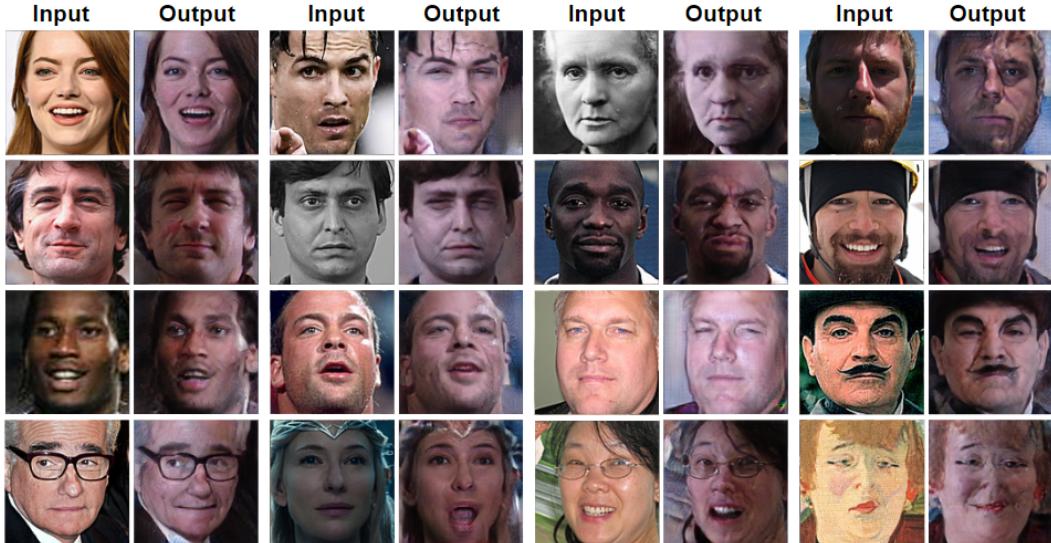


Figure 22: Even more qualitative results demonstrating the generalizability of LEGAN. For each input image, we generate the output image with a randomly selected target lighting and expression. The input images vary in gender, ethnicity, color composition, facial pose, expression, lighting and accessories. All images are $128 \times 128 \times 3$.

6 Conclusion

We propose LEGAN, a GAN framework for performing many-to-many joint manipulation of lighting and expressions of an existing face image without requiring paired training data. Instead of translating the image representations in an entangled feature space like [24], LEGAN estimates transformation maps in the decomposed lighting and expression sub-spaces before combining them to get the desired output image. To enhance the perceptual quality of the synthetic images, we directly integrate a quality estimation model into LEGAN’s pipeline as an auxiliary discriminator. This quality estimation model, built with synthetic face images from different methods [47, 48, 75, 2, 8] and their crowd-sourced naturalness ratings, is trained using a margin based regression loss to capture the subjective nature of human judgement. The usefulness of the the feature disentangling towards synthesis quality is shown by objective comparison [42, 103, 93] of LEGAN to the other GAN models. The same experiments also highlight the usefulness of the proposed quality estimator in LEGAN and (StarGAN w/ L_{qual}) — specifically comparing the latter to vanilla StarGAN. Even when compared with the one-to-one mapping model pix2pix [44], trained to do a single manipulation task using paired data, LEGAN generates a better score for majority of the metrics (Table 8).

As a potential application, we use LEGAN as training data augmenter for face verification and facial expression classification tasks on the IJB-B [95] and AffectNet [61] datasets respectively (Tables 9 and 10). An improvement in the verification score suggests LEGAN can enhance the intra-class variance while preserving subject identity. The boost in expression recognition performance validates the realism of the LEGAN generated facial expressions. The output quality, however, can be further improved when translating from an intense expression to another. We plan to address this by - (1) using attention masks in our encoder modules, and (2) building translation pathways of facial action units [86] while going from one expression to another. Another future goal is to incorporate a temporal component in LEGAN for synthesizing a sequence of coherent frames.

7 Broader Impact

Face image datasets that are collected by scraping the internet have an inherent imbalance in terms of representations from minority communities, extreme pose, lighting and expressions [20, 57, 61]. Using such data to train models translates the bias existing in the datasets on to the model’s predictions. On the other hand, large scale data collection protocols which take into account demographic distributions, are not only resource and time consuming but may not always be possible due to inclement weather or environmental conditions (e.g. a pandemic). In this scenario, having the ability to generate natural looking synthetic data, as LEGAN does, makes the process faster and cheaper

and can be used to mitigate inherent data bias in training. By addressing the issues of data bias and sparsity, LEGAN can improve the generalizability of downstream models trained with the synthesized data.

As is true for any face synthesis work, both LEGAN and our quality estimation model can be used for malicious purposes like generating better looking DeepFakes to manipulate image or video content that can lead to the spread of misinformation. Researchers, while developing face synthesis methods, should also be cognizant of the ethical implications their models might have if the underlying training data is biased or flawed. For example, it can be very simple to overwhelm a downstream classification model with natural looking synthetic images belonging to only particular demographic sub-groups that are abundant in popular face image datasets of celebrities like CelebA [55]. Synthesizing artificial images from this skewed data will not only amplify the underlying bias [71] in classification models, but can have major societal implications depending on its applications [63].

References

- [1] Deepfake Detection Challenge:. <https://www.kaggle.com/c/deepfake-detection-challenge>. 3
- [2] DeepFake FaceSwap:. <https://faceswap.dev/>. 2, 3, 5, 6, 17
- [3] FaceSwap Repository:. <https://github.com/deepfakes/faceswap>. 3
- [4] Pre-filtered StyleGAN Images:. <https://generated.photos/?ref=producthunt>. 4, 5
- [5] M. Alcorn, Q. Li, Z. Gong, C. Wang, L. Mai, W-S. Ku, and A. Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *CVPR*, 2019. 1
- [6] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 9
- [7] S. Banerjee, J. Bernhard, W. Scheirer, K. Bowyer, and P. Flynn. Srefi: Synthesis of realistic example face images. In *IJCB*, 2017. 1, 3
- [8] S. Banerjee, W. Scheirer, K. Bowyer, and P. Flynn. Fast face image synthesis with minimal training. In *WACV*, 2019. Dataset available here: <https://cvrl.nd.edu/projects/data/>. 1, 2, 3, 4, 5, 6, 17
- [9] S. Banerjee, W. Scheirer, K. Bowyer, and P. Flynn. On hallucinating context and background pixels from a face mask using multi-scale gans. In *WACV*, 2020. 2
- [10] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa. The do's and don'ts for cnn-based face verification. *ICCV Workshops*, 2017. 1
- [11] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. In *CVPR*, 2018. 2
- [12] S. Barratt and R. Sharma. A note on the inception score. *arXiv:1801.01973*. 3
- [13] S. Beery, Y. Liu, D. Morris, J. Piavis, A. Kapoor, M. Meister, N. Joshi, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *WACV*, 2020. 1
- [14] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv:1703.10717*. 1
- [15] J.R. Beveridge, D.S. Bolme, B.A. Draper, G.H. Givens, Y.M. Lui, and P.J. Phillips. Quantifying how lighting and focus affect face recognition performance. In *CVPR Workshops*, 2010. 1
- [16] D. Bitouk, N. Kumar, S. Dhillon, S. Belhumeur, and S. K. Nayar. Face swapping: Automatically replacing faces in photographs. *SIGGRAPH*, 2005. 1, 2
- [17] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003. 2

- [18] A. Borji. Pros and cons of gan evaluation measures. *CVIU*, 179(3):41–65, 2019. [2](#), [3](#)
- [19] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. [4](#), [9](#), [10](#)
- [20] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency*, 2018. [17](#)
- [21] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognizing faces across pose and age. In *arXiv:1710.08092*. [1](#), [3](#), [10](#), [11](#), [12](#), [15](#)
- [22] H. Chang, J. Lu, F. Yu, and A. Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *CVPR*, 2018. [2](#)
- [23] J. Chen, G. Su, J. He, and S. Ben. Relighting using locally constrained global optimization. In *ECCV*, 2010. [2](#)
- [24] Y. Choi, M. Choi, M. Kim, J-W. Ha, S. Kim, and J. Chool. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. [1](#), [2](#), [11](#), [12](#), [14](#), [17](#)
- [25] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Face synthesis from facial identity features. In *CVPR*, 2017. [2](#)
- [26] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. [3](#)
- [27] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and Fei-Fei. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. [3](#)
- [28] X. Di and V.M. Patel. Facial synthesis from visual attributes via sketch using multi-scale generators. *TBIOM*, 2019. [2](#)
- [29] I.O. Ertugrul, J.F. Cohn, L.A. Jeni, Z. Zhang, L. Yin, and Q. Ji. Cross-domain au detection: Domains, learning approaches, and measures. In *FG*, 2019. [14](#)
- [30] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92, 1973. [2](#)
- [31] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He. High fidelity face manipulation with extreme pose and expression. *arXiv:1903.12003*. [2](#)
- [32] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 2019. [2](#)
- [33] Z. Geng, C. Cao, and S. Tulyakov. 3d guided fine-grained face manipulation. In *CVPR*, 2019. [2](#)
- [34] 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 *NeurIPS*, 2014. [1](#), [2](#)
- [35] K. Grm, V. Štruc, A. Artiges, M. Caron, and H.K. Ekenel. Strengths and weaknesses of deep learning models for face recognition against image degradations. *IET Biometrics*, 7(1):81–89, 2018. [1](#)
- [36] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing.*, 28(5):807–813, 2010. [10](#), [11](#), [12](#), [14](#), [15](#)
- [37] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017. [9](#)
- [38] X. Han, Y. Liu, H. Yang, G. Xing, and Y. Zhang. Normalization of face illumination with photorealistic texture via deep image prior synthesis. *Neurocomputing*, 2020. [3](#)
- [39] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015. [1](#)

- [40] K. He and X. Xue. Facial landmark localization by part-aware deep convolutional network. In *PCM*, 2016. 4
- [41] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016. 3, 7, 10, 11, 12, 15
- [42] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 2, 3, 10, 11, 12, 15, 17
- [43] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *ICCV*, 2017. 2
- [44] P. Isola, J-Y. Zhu, T. Zhou, and A.A. Efros. Image-to-image translation with conditional adversarial nets. In *CVPR*, 2017. 2, 7, 11, 12, 14, 17
- [45] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. *ICCV*, 2017. 2
- [46] J. Johnson, A. Alahi, and F-F. Li. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 9
- [47] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018. 1, 2, 3, 5, 6, 17
- [48] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *arXiv:1812.04948*, 2018. 1, 2, 3, 4, 5, 6, 11, 17
- [49] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. *CoRR*, arXiv:1912.04958, 2019. 12
- [50] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016. 1
- [51] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 10
- [52] Y.A. Kolchinski, S. Zhou, S. Zhao, G. Mitchell, and S. Ermon. Approximating human judgment of generated image quality. *arXiv:1912.12121*. 2, 3
- [53] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1
- [54] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *CVPR Workshops*, 2015. 4
- [55] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 18
- [56] I. Masi, T. Hassner, A. T. Tran, and G. Medioni. Rapid synthesis of massive face sets for improved face recognition. *FG*, 2017. 1, 14
- [57] I. Masi, A-T. Tran, T. Hassner, G. Sahin, and G. Medioni. Face-specific data augmentation for unconstrained face recognition. *IJCV*, 2019. 17
- [58] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, 2016. 1, 2
- [59] A.L. Mass, A.Y. Hannun, and A.Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 7
- [60] A. Meka, C. Haene, R. Pandey, M. Zollhoefer, S. Fanello, G. Fyffe, A. Kowdle, X. Yu, J. Busch, J. Dourgarian, P. Denny, S. Bouaziz, P. Lincoln, M. Whalen, G. Harvey, J. Taylor, S. Izadi, A. Tagliasacchi, P. Debevec, C. Theobalt, J. Valentin, and C. Rhemann. Deep reflectance fields - high-quality facial reflectance field inference from color gradient illumination. In *SIGGRAPH*, 2019. 3

- [61] A. Mollahosseini, B. Hasani, and M.H. Mahoor. Affectnet: A new database for facial expression, valence, and arousal computation in the wild. *IEEE Trans. on Affective Computing*, 2017. 14, 17
- [62] S. Mosaddegh, L. Simon, and F. Jurie. Photorealistic face de-identification by aggregating donors' face components. In *ACCV*, 2014. 2
- [63] M. Mulshine. A major flaw in google's algorithm allegedly tagged two black people's faces with the word 'gorillas'. 2015. <https://www.businessinsider.com/google-tags-black-people-as-gorillas-2015-7>. 18
- [64] A. Nech and I. Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *CVPR*, 2017. 1
- [65] T. Nestmeyer, J-F. Lalonde, I. Matthews, and A. Lehrmann. Learning physics-guided face relighting under directional light. In *CVPR*, 2020. 3
- [66] Y. Nirkin, Y. Keller, and T. Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, 2019. 2
- [67] T-H. Oh, T. Dekel, C. Kim, I. Mosseri, W.T. Freeman, M. Rubinstein, and W. Matusik. Speech2face: Learning the face behind a voice. In *CVPR*, 2019. 2
- [68] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 1
- [69] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weiment. An introduction to the good, the bad, and the ugly face recognition challenge problem. In *FG*, 2011. 1
- [70] S.M. Pizer and et al. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39:355–368, 1987. 2
- [71] V.U. Prabhu, D.A. Yap, A. Wang, and J. Whaley. Covering up bias in celeba-like datasets with markov blankets: A post-hoc cure for attribute prior avoidance. In *ICML Workshops*, 2019. 18
- [72] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *CVPR*, 2018. 2, 3
- [73] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, 2018. 2, 10
- [74] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2, 7
- [75] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. Available here: <https://github.com/ondyari/FaceForensics>. 2, 3, 5, 6, 17
- [76] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 2, 3
- [77] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 7
- [78] E. Sanchez and M. Valstar. A recurrent cycle consistency loss for progressive face-to-face synthesis. In *FG*, 2020. 15
- [79] S. Sengupta, A. Kanazawa, C. Castillo, and D. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. In *CVPR*, 2018. 3
- [80] W. Shi, J. Caballero, F. Huszar, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 7, 10

- [81] Z. Shu, S. Hadap, E. Shechtman, K. Sunkavalli, S. Paris, and D. Samaras. Portrait lighting transfer using a mass transport approach. In *SIGGRAPH*, 2017. 3
- [82] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 3
- [83] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 2
- [84] T. Sun, J.T. Barron, Y-T. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. Debevac, and R. Ramamoorthi. Single image portrait relighting. In *SIGGRAPH*, 2019. 3
- [85] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3
- [86] Y. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *TPAMI*, 23(2):97–115, 2001. 16, 17
- [87] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. *CVPR*, 2017. 2
- [88] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016. 7
- [89] P. Upchurch, J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Weinberger. Deep feature interpolation for image content changes. *CVPR*, 2017. 2
- [90] R. Vemulapalli and A. Agarwala. A compact embedding for facial expression similarity. In *CVPR*, 2019. 1, 14
- [91] R.G. VidalMata, S. Banerjee, and et al. Bridging the gap between computational photography and visual recognition. *arXiv:1901.09482*, 2019. 1
- [92] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. In *TPAMI*, 2009. 2
- [93] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, 2004. 2, 10, 11, 12, 15, 17
- [94] Y. Wen, B. Raj, and R. Singh. Face reconstruction from voice using generative adversarial networks. In *NeurIPS*, 2019. 2
- [95] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother. Iarpa janus benchmark-b face dataset. In *CVPR Workshops*, 2017. 14, 17
- [96] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *TIFS*, 2018. 14
- [97] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv:1505.00853*, 2015. 4, 7, 14
- [98] J. Yang, Z. Zhao, H. Zhang, and Y. Shi. Data augmentation for x-ray prohibited item images using generative adversarial networks. *IEEE Access*, 7:28894–28902, 2019. 1
- [99] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. In *arXiv:1411.7923*. 14
- [100] H. Yu, O.G. Garrod, and P.G. Schyns. Perception-driven facial expression synthesis. *Computers & Graphics*, 36(3), 2012. 2
- [101] M.D. Zeiler, D. Krishnan, G.W. Taylor, and R. Fergus. Deconvolutional networks. In *CVPR*, 2010. 10

- [102] R. Zhang, P. Isola, and A.A. Efros. Colorful image colorization. In *ECCV*, 2016. 3
- [103] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 3, 10, 11, 12, 15, 17
- [104] H. Zhou, S. Hadap, K. Sunkavalli, and D. Jacobs. Deep single-image portrait relighting. In *ICCV*, 2019. 3
- [105] S. Zhou, M.L. Gordon, R. Krishna, A. Narcomey, Fei-Fei. Li, and M.S. Bernstein. Hype: A benchmark for human eye perceptual evaluation of generative models. In *NeurIPS*, 2019. 2, 3
- [106] J.Y. Zhu, T. Park, P. Isola, and A.A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 9, 11, 12, 14
- [107] K. Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics Gems IV*, 1994. 2