sketch2face: Conditional Generative Adversarial Networks for Transforming Face Sketches into Photorealistic Images

Julia Gong *
Stanford University
Stanford, CA
jxgong@stanford.edu

Matthew Mistele *
Stanford University
Stanford, CA
mmistele@stanford.edu

Abstract

In this paper, we present a conditional GAN image translation model for generating realistic human portraits from artist sketches. We modify the existing pix2pix model by introducing four variations of an iterative refinement (IR) model architecture with two generators and one discriminator, as well as a model that incorporates spectral normalization and self-attention into pix2pix. We utilize the CUHK Sketch Database and CUHK ColorFERET Database for training and evaluation. The best-performing model, both qualitatively and quantitatively, uses iterative refinement with L1 and cGAN loss on the first generator and L1 loss on the second generator, likely due to the first-stage sharp image synthesis and second-stage image smoothing. Most failure modes are reasonable and can be attributed to the small dataset size, among other factors. Future steps include masking input images to facial regions, jointly training a superresolution model, and learning a weighted average of the generator outputs.

1. Introduction

Generation of photorealistic images from sketches of human faces has many creative, commercial, and forensic applications. On the creative side, it may help artists turn sketches into photorealistic content. Commercially, it may generate viable alternatives to purchased stock photos for advertisements. Finally, generation of photorealistic color images from forensic sketches of perpetrators may make it easier for witnesses to confirm or deny likeness matches. This precision during a photo lineup could result in fewer innocent people being detained as suspects on the basis of the forensic sketch.

In this project, we tackle the problem of generating color photorealistic images of human faces from corresponding grayscale hand-drawn sketches. The input to our algorithm is an image of a sketch, and we use an iteratively refined conditional GAN to generate an image of the face the sketch represents.

We aggregate and align datasets (CUHK Sketch Database [12] and CUHK ColorFERET [15]) of facial sketches and corresponding facial photos for training and evaluation. For our baseline, we fine-tune a pretrained conditional GAN, pix2pix, on our training set and evaluate it on the test set. To improve on it, we primarily investigate variations of iterative refinement (IR), in which the generator becomes a sequence of two generators, each with its own loss. We also evaluate the success of incorporating spectral normalization [7] and self-attention [14] into the generator and discriminator.

During training, all the models for this task utilize a sketch-photo pair as input, or the tuple (x_s, x_p) where x_s is the grayscale sketch image and x_p is the color ground-truth photo, and output a generated color photo x_g conditioned on x_s. Input images have 3 RGB channels and are resized and cropped to be 256×256 pixels, and output images are the same dimensions. At test time, the model only receives a sketch, and the generated image is evaluated against the corresponding ground truth photo.

2. Related Work

Wang and Tang [11] introduce the task of synthesizing face photos from face sketches from the CUHK Sketch Databse. They demonstrate that remarkably convincing face photos can be synthesized from the sketches using multiscale Markov Random Fields [12].

More recently, Kazemi et al. [6] tackle the unpaired facial sketch to image problem with GANs. They use a variant of Zhu et al's CycleGAN [16], a well-known image-to-image translation baseline model that uses unpaired images. They modify CycleGAN by adding a facial geometry discriminator network, replacing the spatially local PatchGAN used in CycleGAN (and pix2pix [5]) to encourage the network's discriminator to learn higher-level facial features, and thereby enforce global consistency across the generated

image. They also replace cycle consistency loss with perceptual loss, which improves performance. Wang et al. [13] also use a similar facial feature loss in the related facial aging GAN problem to enforce higher-level consistency of the subject's identity from the input to generated image.

On the paired image translation side, Chen and Hays [2] tackle the more general sketch-to-image problem using the Sketchy database for sketch generation in 125 different image categories. They use a conditional GAN that uses Masked Residual Units (MRUs), designed to allow each layer of the network to decide which portions of the feature map from the previous layer to use in its computation. They use both GAN and classification losses. Overall, their findings are that images cannot always be simultaneously photorealistic and faithful to the original sketch, and though their network is diverse, it does not produce realistic images. The failure modes often show over-faithful generated images that closely follow poorly drawn sketch boundaries, which could be attributed to a lack of generalization to canonical features of the desired class because the network has not learned global characteristics of the classes. This is particularly devastating for generating images from facial sketches due to the artistic liberties that sketch artists may take when drawing and image of a person, especially if recalled from memory.

Isola et al. [5] present the pix2pix model, a conditional GAN for general image-to-image translation. One of their applications is the "sketch to shoe" task, which colors in a plausible image given a sketch for a shoe. Their conditional GAN uses a U-Net [9] encoder-decoder structure with skip connections between encoder and decoder layers of the same feature map dimension, and uses PatchGAN for the discriminator, which enforces local consistency in regions of a certain patch size in the output image.

Like Chen and Hays [2] and Isola et al. [5], we choose to pursue the paired image translation method for this task: first, because we have access to aligned sketch-photo pairs, and second, because we see some potential areas of improvement upon the existing paired translation models with regard to global facial feature consistency, especially for the pix2pix framework on our domain-specific (facial sketch to photo) problem.

We also see potential for improvement using techniques from the following three task-independent GAN papers. Zhang et al. [14] use self-attention layers in their GANs and find that they enable the model to use cues from all feature locations, not just local ones. This idea can potentially contribute model expressiveness to our problem space. Hossain et al. [3] show that the use of a generator followed by a sequence of "editors" that iteratively refine the generated sample can lead to model robustness, which we aim to achieve in ours. Miyako et al. [7] introduce a normalization technique called spectral normalization, which they show can effectively stabilize the training of GAN discriminators.

3. Methods

Generative Adversarial Networks (GANs) are generative models that learn a mapping from a random noise vector z to an output image y. In this paper, the model we use is a conditional GAN, which permits observation of an input image x in generating the output image. Conditional GANs therefore learn to map a random noise vector z along with an input image x to an output image y. The conditional GAN consists of a generator network G and discriminator network D. The game-theoretic interpretation of the network formulation is that G tries to generate fake images that are as similar as possible to the real image, y, to fool D, and D tries to be accurate in distinguishing real images from generated images and outputs a likelihood score of the image being real that lies between 0 and 1.

3.1. Baseline Model: Conditional GAN

3.1.1 Objective

The standard conditional GAN objective that the generator minimizes and the discriminator maximizes is

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))]. \quad (1)$$

The term $\mathbb{E}_{x,y}[\log D(x, y)]$ denotes the discriminator's loss from predicting whether the real image is real or fake, while the second term denotes the discriminator's loss from predicting whether a generated image $G(x, z)$ is real or fake. Since the discriminator hopes to achieve as low loss as possible in classifying the real and fake images, the generator in turn maximizes this loss. Therefore, the optimal generator model can be formulated as the optimal solution to the minimax problem $G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D)$.

The conditional GAN also utilizes another loss that takes advantage of the paired dataset to boost the accuracy of the images generated. As noted by [5], L1 distance is favorable to L2 distance since it discourages blurring. The L1 distance loss for target y and generated image $G(x, z)$ is

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{x,y,z}[||y - G(x, z)||_1]. \quad (2)$$

The overall objective for the baseline pix2pix model with L_1 loss weight λ (1 by default, which we use) is therefore

$$\mathcal{L}_{baseline} = \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L_1}(G). \quad (3)$$

3.1.2 Architecture

Our baseline method for this task is a finetuned pretrained model implementation from the paired image-to-image translation conditional GAN model, pix2pix by Isola

et al. [5] and implemented by Zhu et al. on GitHub. The generator is a U-Net [9], a convolutional encoder-decoder model with skip connections between layers of identical feature map size, trained on $256 \times 256 \times 3$ images. The discriminator is a PatchGAN 3-layer classifier introduced by [5]. The pretrained weights used were from the edges2shoes version of pix2pix, which was trained on sketch-photo pairs of colorful shoes. The pretrained model only included generator weights and not discriminator weights, so the model was trained for 110 epochs on top of the pretrained generator weights and discriminator weights using normal initialization until model results began to plateau.

3.2. Conditional GANs with Iterative Refinement

On this particular sketch-to-photo task, the pix2pix model suffers from deficiencies in global consistency across the image and lacks smooth and crisp boundary lines for separating colors from one another on the face. Intuitively, after the initial output image is produced by the model, an additional network that learns to modify the image to make it more realistic would be helpful. To do this, inspired by the concept of iterative refinement, we implement and test four modifications of the pix2pix baseline model.

The iterative refinement (IR) architecture builds off of the conditional GAN framework and uses the same discriminator model as the baseline; however, it instead involves two generator segments that combine to form the generator network. We use U-Net 256 models for both generators. The output of the first generator is subject to a set of losses conditioned on the target image, and its output is fed to the second generator, which then also generates an image subject to another set of losses conditioned on the target image. The hope is for the network to learn to use the second generator to fine-tune the image obtained from the first.

To train this model, we perform transfer learning on top of the baseline's weights for 200 epochs. Specifically, we initialize both G_1 and G_2 to the baseline G weights, and the baseline D weights are used as the initial weights for D.

Both generators of this network can be evaluated with L1 or cGAN losses (or both). We now describe the four variants of the IR architecture that explore these design options.

3.2.1 IR Model with cGAN-Final Loss

The baseline model's objective uses two losses: the cGAN loss and the L1 loss. One important design choice for the IR model is which losses to impose on the two generators. We first try imposing both the cGAN loss and L1 loss on the second generator, but only an L1 loss on the first generator. Both L1 losses share the same λ parameter. We call this objective the 'cGAN-Final' loss. Thus, the objective for generators G_1 and G_2 and discriminator D is

$$\begin{aligned}\mathcal{L}_{cGANFinal} = \mathcal{L}_{cGAN}(G_2, D)+ \\ \lambda\mathcal{L}_{L_1}(G_1) + \lambda\mathcal{L}_{L_1}(G_2).\end{aligned} \tag{4}$$

This model architecture is illustrated in Figure 1. Conceptually, the idea is to let the model decide what to do with the first generator aside from generating images similar to the ground truth, while the final result needs to 'pass the test' of the discriminator.

3.2.2 IR Model and cGAN-Initial Loss

The cGAN-Initial Loss model has the same architecture and L1 loss as that of the cGAN-Final Loss model; however, the first generator receives the cGAN Loss rather than the second generator. Concretely, its objective is

$$\begin{aligned}\mathcal{L}_{cGANFinal} = \mathcal{L}_{cGAN}(G_1, D)+ \\ \lambda\mathcal{L}_{L_1}(G_1) + \lambda\mathcal{L}_{L_1}(G_2).\end{aligned} \tag{5}$$

3.2.3 IR Model and cGAN-Both Loss

The cGAN-Both Loss model places cGAN and L1 losses on both generators, yielding an objective of

$$\begin{aligned}\mathcal{L}_{cGANFinal} = \mathcal{L}_{cGAN}(G_1, D)+ \\ \mathcal{L}_{cGAN}(G_2, D)+ \\ \lambda\mathcal{L}_{L_1}(G_1) + \lambda\mathcal{L}_{L_1}(G_2).\end{aligned} \tag{6}$$

3.2.4 IR Model with cGAN-Final Loss and Grayscale

The previous three IR models allow the network to decide what to learn in the first generator and what refinements to make in the second. However, it is also worth exploring explicitly designed tasks for each generator. Thus, we also create an IR network whose first generator's target image, instead of being the same as that of the second generator, is the grayscale version of the color target. We thus force the first generator to map sketches to grayscale images, which might be an easier task for the network than translating to a color image. The second generator then colorizes and refines the grayscale output of the first generator. Only the cGAN-Final Loss was tried for this model in the interest of having one discriminator model that discriminates over color images only. The overall objective is the same as (4).

3.3. Spectral Normalization and Self-Attention

Our last model incorporates two additional techniques: spectral normalization and self-attention. Both require changes to the network that necessitate training the models from scratch, due to the way the pretrained edges2shoes model was constructed. We train this model for 110 epochs.

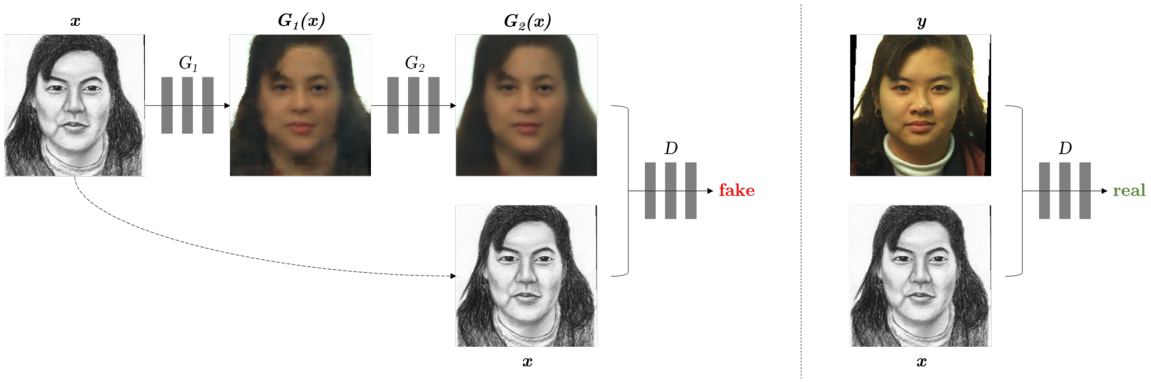

Figure 1. Architecture of iterative refinement model with color images and first loss variant: cGAN-Final Loss. D learns to distinguish between generated and real images, while G_1 and G_2 jointly learn to generate realistic photos from input x.

As defined by Miyako et al. [7], spectral normalization is the process of dividing weight matrices by their largest singular value, in order to make that largest singular value equal 1. This process is tied to what is known as the "Lipschitz constant" of the function, and fixing the constant at 1 bounds the gradients in the discriminator and the generator [7]. We adapt code from this GitHub to add a spectral normalization layer after each convolution layer in our generator and discriminator.

We also investigate the effect of adding a self-attention layer to each generator and discriminator in our IR cGAN, as Zhang et al. [14] do in SAGAN. Our self-attention layer applies affine transformations followed by a ReLU nonlinearity to the input to produce "query", "key", and "value" tensors. We matrix-multiply the query and key tensors, apply a softmax to form an attention distribution over the input features, and matrix-multiply it by the "value" tensor to produce self-attention feature maps. We use a skip-connection to carry forward the original features, and we concatenate these features to the self-attention maps to form the output.

In the "Spectral Normalization and Self-Attention" model, we incorporate self-attention layers after the third main convolution layer in the discriminator, and after the third level deep of recursion in the U-Net of each generator. We adapt Self-Attention code from the SAGAN GitHub for these layers, modifying it to work with the nn.Sequential framework with which the model was set up.

4. Dataset

The dataset we have assembled consists of 1037 sketch-photo pairs that come from two separate datasets: the CUHK ColorFERET Sketch Database [15] ('ColorFERET', 849 pairs) and the CUHK Face Sketch Database [12] ('CUHK', 188 pairs). While we also completely preprocessed the IIIT-D Sketch Database [1] (375 pairs), we did not end up using it in our work because of face alignment issues that were not resolvable within the project time frame. Our train-test split is 788 train-249 test, with a 80-20 split on the ColorFERET dataset and the 88-100 split that the CUHK dataset already designates. Figure 3 shows some example images from the dataset. The ColorFERET database required extensive pre-processing.

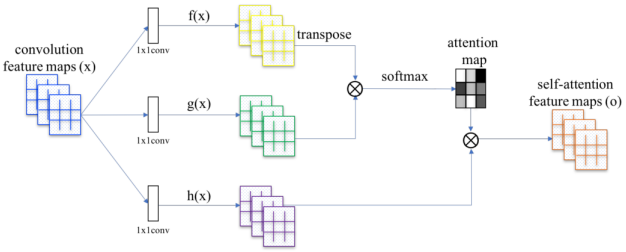

Figure 2. Self-attention mechanism as presented in [14]

We crucially note that these sketch-photo pairs do not use 'reverse-engineered' or generated sketches from unpaired photos; the sketches were drawn manually by sketch artists, and thus contain a higher degree of natural variation in its boundary lines and overall shapes. We make this choice despite the smaller amount of data because we believe it elevates the problem difficulty, as these are more realistic representations of human sketches that force the network to go beyond boundary-matching when filling in the colors of the grayscale sketch.

4.1. Pre-processing

The ColorFERET database came as a deeply nested set of folders containing photographs from two DVDs and two CDs, a corresponding list of photo names that corresponded to sketches, a folder of sketches, and a separate folder of fiducial points for image alignment. The folder of each photographed individual had multiple photographs from various angles, and file paths were not consistent between the

Figure 3. Sample images from dataset, with sketches on first row and color images on second row. The first two pairs are from CUHK and last two are from ColorFERET.

list of file names provided by CUHK and the ColorFERET database (some listed files were also not existent). As a result, pre-processing to extract the correct images included writing scripts for 1) extracting the corresponding color image for each sketch image, 2) applying corrections to these extractions based on an error log file in the dataset, and 3) mapping the true sketch image names to the true color image names.

We also derived the affine transformation required to linearly map three fiducial points to three given locations, wrote a script to do this PIL and cv2, and used the fiducial points provided with ColorFERET along with the consistent and discernable fiducial points of CUHK's cropped images to align the eyes and mouth of each sketch-photo pair in a consistent way across the datasets.

To augment the dataset and promote robustness, we take random crops of the image pairs on each train iteration. The network input is a resized $256 \times 256 \times 3$ image.

5. Results and Discussion

5.1. Hyperparameter Selection

As our model builds on top of the edges2shoes pix2pix model, we use the optimal hyperparameters reported by Isola et al. [5]: learning rate of 0.0002, batch size of 1, $\beta_1 = 0.5$ for the Adam optimizer, normal weight initialization, as well as $\lambda = 1$ for the L1 loss weight. Our focus is on developing the strongest model architectures and objective functions using these pre-selected optimal parameters.

5.2. Qualitative Results and Analysis

To qualitatively evaluate the model, we present some example generated images from running the baseline pix2pix model and the four iterative refinement models on the withheld test set (Table 1).

Visually, the baseline model appears to have learned rough, general facial structure, but not global consistency across facial features, with inconsistent halves of the face or miscolored or color-jittered portions of the face that appear unrealistic. The faces have abrupt and unnatural transitions and some extraneous artifacts, possessing more artistic than photorealistic qualities. However, the model has learned some variation in skin color, as seen across all images, as well as hair, lip, and cheek color. None of the models appear to have learned crisp templates for clothing (as is expected).

Some very interesting patterns emerge in the images from the IR models. First, it appears that all four models perform qualitatively better than the baseline in producing more realistic and plausible images that are also closer to the ground truth image. There are fewer undesired artifacts, and the images tend to be more globally consistent and smoother in terms of facial features, and the image colors have less jitter. They also give a more realistic sense of the 3D structure of the face. The skin tones are also closer to those of the ground truth images.

It's also interesting to note that the model does not simply learn to correlate common masculine or feminine facial features with hair length. For example, it produces consistent masculine features on male subjects with long hair and feminine features for female subjects with short hair. Thus, there is evidence that the model is truly learning mappings from features in the sketches to those in the target images.

Among the IR models, the images produced by generators subject to the cGAN loss tend to have more color jitter and sharp edges similar to the baseline model, though not as salient. On the other hand, images produced by generators subject to only the L1 loss tend to have smoother, a bit blurrier, yet more even and visually realistic faces. It appears that in models where only one generator has the cGAN loss, the generator with only the L1 loss acts as a sort of 'smoothing' filter to the image. The human judgment that these images are more realistic and closer to the target image likely results from the blurriness, since there are fewer jarring, unnatural color variations and edges in the image. Their template-like quality for salient features of the face makes them more plausible.

Intuitively, the L1 loss is a pixel-level loss, and here, it appears to act as a smoother that reduces pixel variation, trading off crisper boundaries for more averaged and 'safe' guesses for plausible pixel values. Smudging boundaries gives the image softer penalties because pixel values do not have extreme transitions. Smoothing lowers loss, since if a sharp edge is off by even just a pixel, the model is greatly penalized. When the cGAN loss is added in addition to the L1 loss, the cGAN loss dominates and pushes the model toward making crisper boundaries that might have a chance at fooling the discriminator, since the discriminator can easily rely on recognizing blurring and soft pixel transitions to distinguish between real and fake images. An important parameter that can better balance this tradeoff is the λ in equations (3, 4, 5, 6), which we did not have time to tune but would be a good direction to explore.

As for the cGAN-Final Loss model with spectral normal-

ization and self-attention, images produced by the first generator are comparable to those of the first generators in other models; however, the images from the second generator are significantly more pixellated in some places than the others. Even the training image outputs have pixellated patches, which suggests more training time may be needed to fit the extra parameters added by spectral normalization and self-attention. Unlike the other models, this model does not have the starting point of a pretrained model, and it shows: either the second generator or the discriminator needs more training to do better. Another possible cause of the pixellation is extreme gradients, but the spectral normalization makes this less likely.

Though there was insufficient time to conduct a thorough user study, some casual surveys of a handful of subjects resulted in the assessment that the cGAN-Initial second generator (2b) and cGAN-Both first generator (3a) images are the most realistic and closest to the ground truth image.

5.3. Quantitative Results and Analysis

We choose five evaluation metrics for evaluating model performance, the values for all of which are arithmetic averages (means) across outputs of the test set: L1 and L2 distance, SSIM (Structural Similarity Index), and FaceNet [10] Embedding distance.

L1 and L2 distance are standard for pixel-level image comparison, and SSIM is standard for evaluating structural similarity in images that evaluates the quality of a processed image from a true image, which aligns with this task. Finally, little-seen in the literature is an evaluation metric based on facial feature embedding distances (though identity verification rates have been used, as in [6]), which we introduce here as the comparison of FaceNet embedding distances using the pretrained FaceNet model from here.

See Table 2 for these values for all models and their output images. All of the iterative refinement models outperform the baseline by a large margin. The output of the cGAN-Initial model's generator 2 has the strongest performance among the L1 and SSIM metrics, though it is about average for L2 distance, where the cGAN-Final generator outputs shine. However, it has a much higher SSIM than any other model by a significant margin, which shows that the structural similarity of the faces produced by the cGAN-Initial generator 2 are closest to those of the target.

However, the story is very different for the FaceNet embedding distance metrics; the best models for the other three metrics and human evaluation have the worst FaceNet embedding metrics, and the worst models for the other metrics have the best FaceNet embedding metrics. Seeing as the other three metrics coincide much more closely with human evaluation, we are inclined to believe that these particular FaceNet embeddings don't necessarily capture well, at least in this context, the concept of facial 'realness' or similarity that humans perceive. Therefore, overall, the cGAN-Initial model appears to perform the best out of the four iterative refinement models, especially with the output of generator 2. We hypothesize that this model works well because imposing both the cGAN and L1 losses on the first generator creates an image with crisp boundaries that can fool the discriminator; however, it exhibits sharp color transitions and jitter that doesn't appear natural to the human eye, which is then smoothed out by the second generator that is only subject to L1 loss. The cGAN-Both model also has reasonable results, but the images appear less natural because less smoothing occurs when both generators use the cGAN loss.

Finally, to highlight some takeaways from the evaluation metrics, the L1 and L2 distance operate on the pixel level, which is a good heuristic, but does not necessarily capture the way humans compare image similarity. On the other hand, the SSIM and FaceNet metrics explore higher-level feature comparison. However, despite the fact that SSIM is not specific to facial features and L1 and L2 distance are lower-level metrics, these metrics are more consistent with human evaluation, while the FaceNet embeddings yield results that are highly inconsistent with human evaluation. This suggests that there is still room for developing better evaluation metrics that capture the human-like comparison of faces on the level of facial features, an exciting direction for future work.

5.4. Failure Modes and Shortcomings

Table 3 highlights examples of the main types of failure modes observed in the test set outputs. Starting from the first row, the main deficiencies across all models with varying degrees of severity are: failure to reproduce less common hairstyles that involve asymmetry or obstruction of the face (1), incomplete or malformed facial contraptions, such as glasses shown (2), odd or unfaithful coloration of headgear such as headbands and hats (3), 'smoothing away' of details such as earrings, inferring drastically incorrect (usually lighter) skin tone and hair color (4, 5), and unrealistic residual artifacts on less common facial or cranial shapes (6). However, we do note that the model does not suffer from mode collapse (producing low-diversity samples or the same sample repeatedly given different inputs).

On a high level, the likely reason for most of these failure modes is the small dataset size. The model successfully learns the general template for a human face with various key features, but simply hasn't seen enough examples of variations in earrings, glasses, and different hairstyles and skin tones to be able to faithfully reproduce them from a detailed sketch. With larger paired datasets with greater diversity in the types of detail found in these images, we suspect the model will perform much better.

There are also more nuanced factors that might contribute to these errors. First, some images in the dataset have

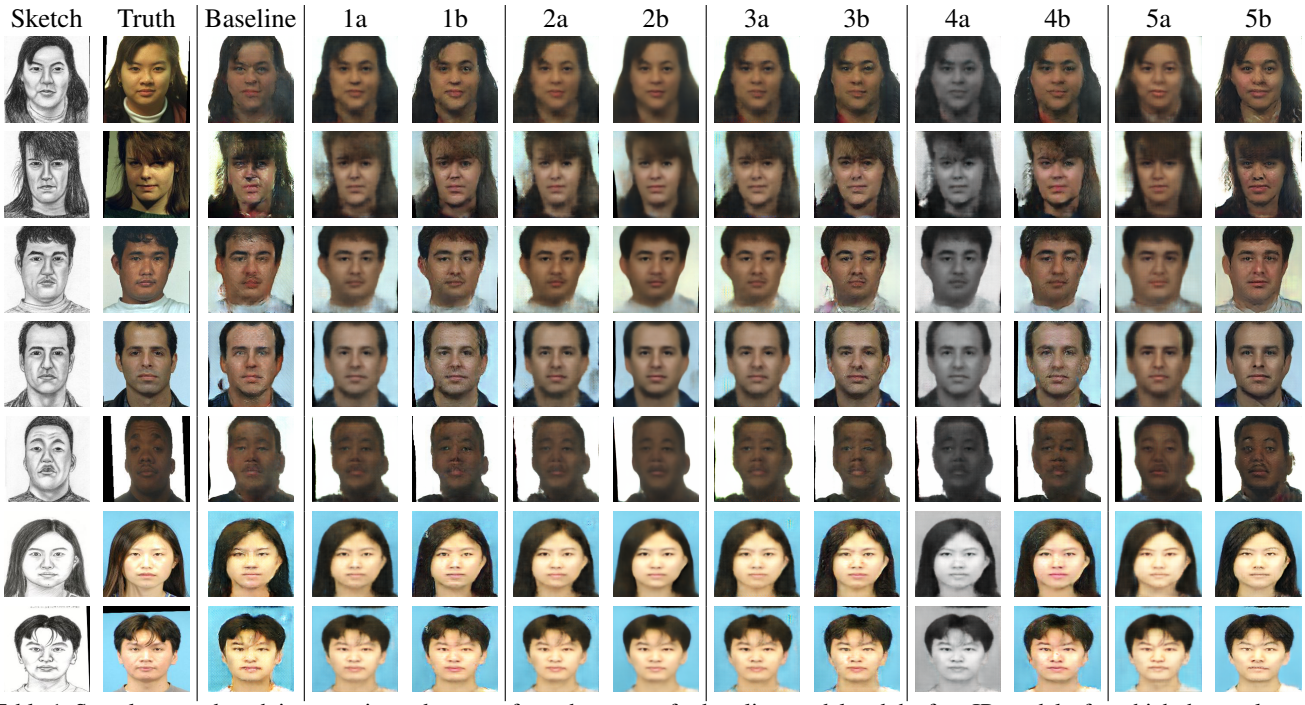

Table 1. Sample ground-truth image pairs and outputs from the test set for baseline model and the four IR models, for which the results are presented in the order cGAN-Final (1), cGAN-Initial (2), cGAN-Both (3), cGAN-Final with grayscale (4), and cGAN-Final with Spectral Normalization and Self-Attention (5). 'a' refers to the output of the first generator, and 'b' to the second.

	L1	**L2**	**SSIM**	**FaceNet L1**	**FaceNet L2**
Baseline	35.245	56.928	0.487	43.147	1.999082
cGAN-Final generator 1	32.579	**53.332**	0.579	43.157	1.999086
cGAN-Final generator 2	33.647	55.367	0.524	43.098	1.999075
cGAN-Initial generator 1	32.523	53.713	0.581	43.171	1.999085
cGAN-Initial generator 2	**32.273**	53.794	**0.608**	43.220	1.999104
cGAN-Both generator 1	32.705	53.490	0.564	43.137	1.999084
cGAN-Both generator 2	33.564	55.290	0.529	43.102	1.999073
cGAN-Final gray	33.409	55.135	0.526	**43.088**	**1.999072**
cGAN-Final w/ spectral norm & self-attn, gen. 1	33.721	54.595	0.571	43.164	1.999092
cGAN-Final w/ spectral norm & self-attn, gen. 2	34.811	57.038	0.530	43.090	1.999072

Table 2. Average L1 distance, L2 distance, SSIM, and FaceNet [10] embedding L1 and L2 distance metrics across the 294 test set outputs for all models. Best scores are bolded (SSIM is the only one for which a higher score is better). Note that FaceNet L2 distance has variances on the order of 1e-9, so more significant digits are displayed.

poor lighting, such as the third subject in Table 3, resulting in a darker and flatter perception of the face. Since the model is trying to learn distinguishing facial features, the bad lighting reduces some of this structural 3D information. Thus, the corresponding outputs appear more flat and facial features sometimes malformed. In addition, as mentioned previously, the sketches are hand-drawn by artists and not synthesized; therefore, some sketches are not extremely accurate and faithful to the target image, as in the first subject in Table 3. This makes it more difficult for the model to properly learn the mappings from sketch outlines to facial features, and given this handicap, we believe the model performs reasonably well. Unsurprisingly, the more successful images in Table 1 tend to have sketch-image pairs that have highly similar details and relative positioning.

Another concern upon observing the outputs is that the model clearly learns the background colors of the images, since the background colors are almost perfectly reproduced in Table 1 and often in 3. The background color reproduction suggests that the model can consistently determine which dataset a given sketch belongs to, and may even condition the image generation on this knowledge

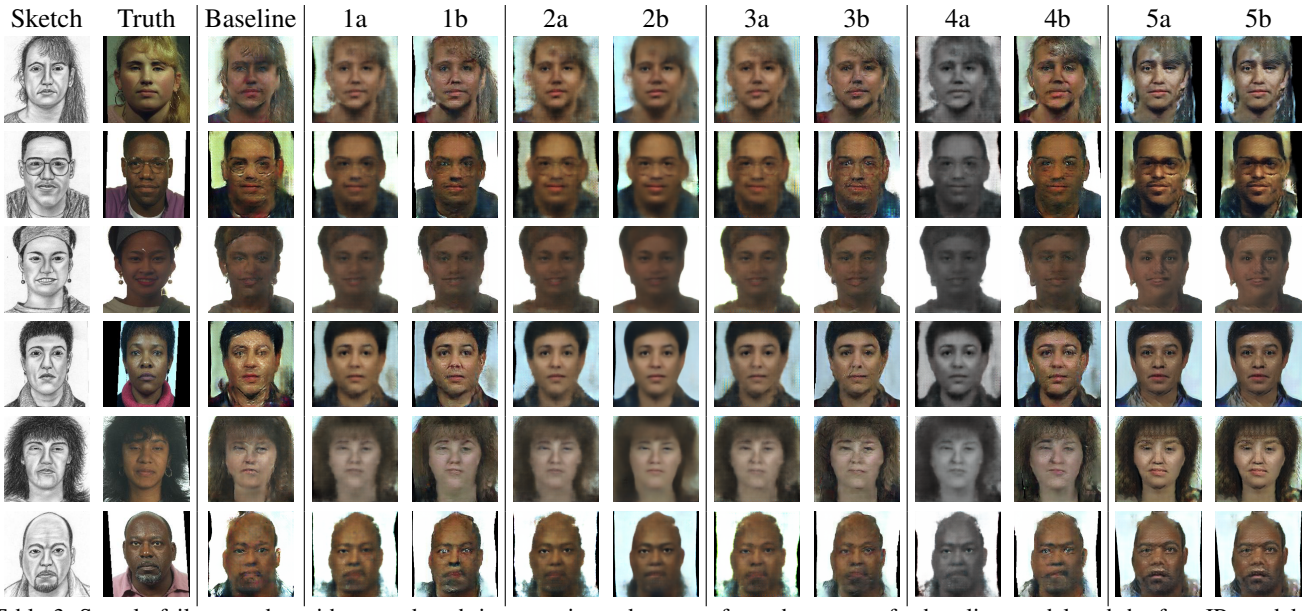

Table 3. Sample failure modes with ground-truth image pairs and outputs from the test set for baseline model and the four IR models, for which the results are presented in the order cGAN-Final (1), cGAN-Initial (2), cGAN-Both (3), cGAN-Final with grayscale (4), and cGAN-Final with Spectral Normalization and Self-Attention (5). 'a' refers to the output of the first generator, and 'b' to the second.

since the background colors of the two datasets are so well-reproduced and so distinct. This potential 'memorization' yields some concerns, such as the robustness of the model in generalizing to sketches of similar quality but drawn by other artists. The model may operate on a narrow domain in terms of sketch style since it may have effectively learned a mapping from artist sketch style to image features. Moreover, the model should not even be penalized for incorrect background or clothing—only the facial features need to be learned. Thus, future work might incorporate a masking technique to only train the model on face pixels.

6. Conclusion and Future Work

In this paper, we present a new conditional GAN image translation model for the task of generating images of human faces from artist sketches. We build our work off of the pix2pix model [5] and the concept of iterative refinement (IR) and train our models by fine-tuning the edges2shoes pix2pix model. We present four formulations of the iterative refinement cGAN and also test spectral normalization and self-attention. The model performing best overall qualitatively and quantitatively is the iterative refinement cGAN-Both model that uses a generator with L1 and cGAN loss that feeds into a second generator that uses only L1 loss. We attribute its success to the two-step mechanism of this model: the output of the first generator has crisp boundaries and resembles the target in a close but "choppy" fashion, and the second generator smooths it in a way that makes the image more plausible.

We see many avenues for future work on this task. First, gathering more data will likely provide performance boosts. Including more paired datasets like those used here ([12] and [15]) and [1] that have greater ethnic diversity and representation of varied facial features, hair, and head garments will result in more detailed and realistic images.

To improve on our current model's design, the network inputs can be masked to omit background and clothing pixels to only train the model on face pixels. This will prevent the model from learning arbitrary correlations between facial features and background color, for example. We'd also like to try non-RGB color spaces, such as HSL, HSV, and CMYK, to determine which is most effective for this task.

Since the best performing networks have blurrier output images, it would be interesting to perform iterative refinement on these smooth output images using a superresolution model to yield crisp yet realistic images. We also consider outputting a learned weighted average of the outputs of the two generators, to reconcile image detail with smoothness.

There is also room for improvement on the cGAN-Final model with the grayscale intermediate generator. The second half of the network is a colorization network; we could first pretrain this network using a much larger face colorization dataset, such as Labeled Faces in the Wild (LFW) [4], before training this network end-to-end.

Finally, inspired by the skin tone failure modes from Table 3, an interesting research direction may incorporate channel autoencoders [8] or another mechanism that allows manipulation of the latent space of the cGAN, perhaps controlling skin tone, eye color, hair color, and other variables to correct for incorrect inferences made by the model.

7. Contributions and Acknowledgements

We have thoroughly enjoyed working on this project, and hope that you may share the same excitement in reading our findings.

J.G. cleaned and preprocessed the raw datasets to extract all the sketch-photo pairs. M.M. wrote the code to perform affine transformation for aligning faces across the dataset. M.M. modified the data pipeline code from pix2pix to pull from multiple dataset folders and to pair up CUHK sketch images and corresponding face images. J.G. did the same for the ColorFERET dataset. M.M. extended pix2pix's preprocessing code to augment the dataset by randomly cropping both images of a training pair in the same way. J.G. trained and evaluated the baseline model, and wrote the code for, trained, and evaluated the four iterative refinement models by changing the architecture and loss functions of the baseline. M.M. wrote the code for, trained, and evaluated the spectral normalization and self-attention model, adapting GitHub implementations for spectral normalization and self-attention. J.G. implemented the FaceNet embedding evaluation metrics by writing a wrapper around the FaceNet PyTorch code to make it compatible with the dataset. M.M. implemented the L1 and L2 evaluation metrics and used the skimage package for SSIM evaluation.

For the milestone report, M.M. generated figures and J.G. wrote the content. For the final report, M.M. extended Introduction and Related Work, and wrote on the spectral normalization and self-attention model and analyzed its results; all other content and figures were written by J.G.

We would like to thank Yannick Hold-Geoffroy and Cynthia Lu of Adobe Research and Lecturer Justin Johnson of CS 231n for discussing ideas with us about the project direction and potential network designs and evaluation methods. We also want to express sincere gratitude to the CS 231n course staff for their continued encouragement and for putting together a remarkable computer vision course.

References

[1] H. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. Memetically optimized mcwld for matching sketches with digital face images. In *IEEE Transactions on Information Forensics and Security*, 2012.

[2] W. Chen and J. Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. *CoRR*, abs/1801.02753, 2018.

[3] S. Hossain, K. Jamali, Y. Li, and F. Rudzicz. Chaingan: A sequential approach to gans, 2018.

[4] G. B. Huang, M. Mattar, T. Berg, and E. Learned-miller. E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 2007.

[5] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[6] H. Kazemi, F. Taherkhani, and N. M. Nasrabadi. Unsupervised facial geometry learning for sketch to photo synthesis. *CoRR*, abs/1810.05361, 2018.

[7] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR 2018*, 02 2018.

[8] T. J. O'Shea, K. Karra, and T. C. Clancy. Learning to communicate: Channel auto-encoders, domain specific regularizers, and attention. *CoRR*, abs/1608.06409, 2016.

[9] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[10] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[11] X. Tang and X. Wang. Face sketch synthesis and recognition. pages 687–694 vol.1, 11 2003.

[12] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 31, 2009.

[13] Z. Wang, X. Tang, W. Luo, and S. Gao. Face aging with identity-preserved conditional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks, 2018.

[15] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[16] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.