Attentive Conditional Channel-Recurrent Autoencoding for Attribute-Conditioned Face Synthesis

Wenling Shang
UvA-Bosch DELTA Lab
wshang@uva.nl

Kihyuk Sohn
NEC Labs
ksohn@nec-labs.com

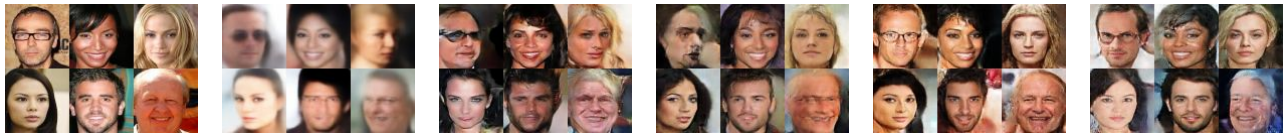


Figure 1: Generations using attributes from ground truth images. Comparison shows our attentive conditional channel-recurrent VAE-GAN's superiority in attribute-conditioned face synthesis.

Abstract

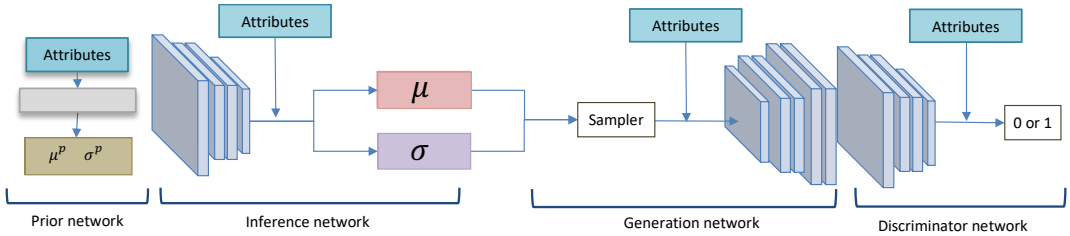
Attribute-conditioned face synthesis has many useful applications, such as to aid the identification of a suspect or a missing person. Building on top of a conditional version of VAE-GAN, we augment the pathways connecting the latent space with channel-recurrent architecture, in order to provide not only improved generation qualities but also interpretable high-level features. In particular, to better achieve the latter, we further propose an attention mechanism over each attribute to indicate the specific latent subset responsible for its modulation. Thanks to the latent semantics formed via the channel-recurreny, we envision a tool that takes the desired attributes as inputs and then performs a 2-stage general-to-specific generation of diverse and realistic faces. Lastly, we incorporate the progressive-growth training scheme to the inference, generation and discriminator networks of our models to facilitate higher resolution outputs. Evaluations are performed through both qualitative visual examination and quantitative metrics, namely inception scores, human preferences, and attribute classification accuracy.

1. Introduction

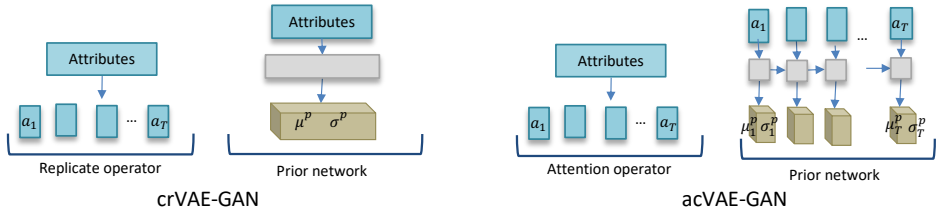
Assaults, burglaries and thefts are among the top common crimes taking place in the United States [11]. The post-trauma memories of the witnesses become susceptible to even a short period of delay, due to factors such as stress and potential contamination from other events. It is shown that the recall of a suspect's face from a witness is the most accurate within 3-4 hours after the incidence and drastically declines after 2 days [12]. However, traditional sketching methods such as Forensic artists or composition software [1] both require a certain level of human expertise, where additionally the former takes a long period of time to complete and the latter can be less descriptive [26]. To this end, we strive for a data-driven tool to synthesize realistic and diverse face candidates basing on attribute information that returns instantaneous feedback to aid the identification of e.g. a crime suspect.

Deep conditional generative modeling is a natural approach to such attribute-conditioned face synthesis application. There has been tremendous recent progress in constructing such models for a variety of tasks. Applications related to generating face images have been especially well-studied, with techniques mainly derived from the 3 most dominating generative model frameworks, namely autoregressive models [38], generative adversarial networks (GANs) [15] and variational autoencoders (VAEs) [24].

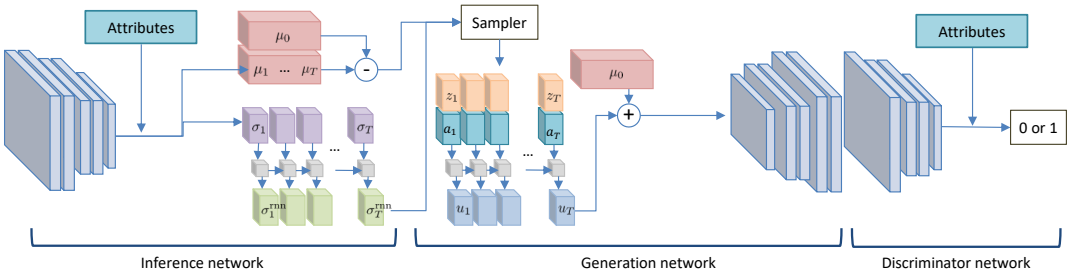
Conditional autoregressive models have been proposed to synthesize images from conditional variables such as text, key points and the initial frame of a video [40]. However, few works have applied this method on faces, likely because the pixel level autoregression is computational costly and slow in inference time without complex parallelization. Conditional GANs, on the other hand, have been more widely utilized on faces. In particular, [14] generates attribute-conditioned faces from scratch but the outputs (Figure 1) can be flawed with artifacts. Conditional VAE [25] (cVAE)—an extension of VAE [24]—is a conditional directed graphical model to approximate conditional data distribution through maximizing its variational lower bound. In [48], attributes and background masks are used



(a) Prior, inference, generation and discrminator networks for cVAE-GAN



(b) Manipulation of attributes and prior for crVAE-GAN and acVAE-GAN



(c) Inference, generation and disriminator networks for cVAE-GAN and acVAE-GAN

Figure 2: Illustration of models used in our work.

as conditional variables in their cVAE formulation to generate faces. Although the reconstruction component in the variational lower bound gives the advantage of a more comprehensive input space coverage opposed to cGAN, cVAE suffers from blurry generation when optimized in combination with a highly restricted KL regularization on the approximated posterior (Figure 1).

To alleviate the blurriness in generation while maintaining the probabilistic latent space, [28] augments a VAE with an adversarial loss, arriving at VAE-GAN. As a follow-up, [4] formulates a conditional version of VAE-GAN, cVAE-GAN for face synthesis, where the conditional variables are identity labels, i.e. one-hot categorical vectors. However, this model is not directly applicable for our application as it would require 2^N-dimensional conditional variable to represent all possible attribute combinations, where N is the number of attributes. Therefore, in this work, we establish a cVAE-GAN baseline specifically targeting at generations out of high-dimensional attribute information (Figure 1), where the attribute condition is embedded in a N-dim binary vectors instead one-hot vectors.

On top of our baseline, we integrate the channel-recurrent architecture [45] where the latent space is divided into consecutive and non-overlapping blocks that are connected via a recurrent module, leading to conditional channel-recurrent VAE-GAN (crVAE-GAN). The benefit of doing so is twofold. Firstly, by introducing a more complex architecture to maneuver the pathways into and out of the latent space characterized by simple diagonal Gaussians, the image generation quality is substantially enhanced (Figure 1). Secondly, the channel-recurrent architecture captures the high-level information in a course-to-fine manner, allowing more explainable latent features, which inspires us to propose our final model, the attentive conditional channel-recurrent VAE-GAN (acVAE-GAN). Our acVAE-GAN learns attention over the attribute vectors so that each attribute attends a specific latent block. In addition to being responsible for different attributes, the channel-recurrent layout also assigns different content for each latent block to modulate: some block predominately controls the global content whereas others focus more on finetuning locally. This unique property of channel-recurrency enables us to envision an application tool to performs a 2-stage general-to-specific conditional face generation.

Lastly, we incorporate progressive-growth training [21] to the cVAE-GAN framework to facilitate higher resolution outputs.

The merits of our work are summarized as follows:

- Construct the cVAE-GAN baseline for attribute-conditioned face synthesis.
- Improve the generation quality of cVAE-GAN by integrating the channel-recurrent architecture, arriving at

crVAE-GAN.

- Towards more interpretable models, learn an attention mask over attribute vectors so that each attribute "chooses" to be modulated by a specific latent block.
- Implement progressive-growth to our models to increase generation resolution.
- Envision a tool that performs 2-stage general-to-specific attribute-conditioned face generations.

2. Related Works

In the regime of deep generative modeling, the most significant recent progress comes from autoregressive models, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). Originally framed as unsupervised learning, all of the three themes have also evolved to model conditional distributions with side information.

Algorithms such as Pixel RNN and Pixel CNN [37, 38, 43] explicitly model the pixel space distribution in an autogressive manner. Conditional versions [13, 40] have been developed to generate a scene from text, a video from an initial frame, a segmentation mask from an image, etc. Although granting high quality outputs and an exact log-likelihood, these methods do not learn a latent representation and demand heavy GPU parallelization for training.

GAN [15] is another mainstream method to render sharp images where a generator attempts to fool a discriminator with generations resembling examples from the input space. Conditional GANs (cGANs) [10, 20, 34] introduce conditional information to the generator and the outputs are expected to not only inmitate the inputs but also obey the conditions. Consequently, the discriminator must learn to distinguish correctly-matched real-condition pairs from both fake-condition paris and wrongly-matched pairs [39]. Adversarial training suffer from several notable drawbacks, such as limited input space coverage [3, 46, 47], generation artifacts and lack of probabilistic latent representations. The training of GANs are also known to be unstable, albeit there have been many works attempting to address this, for both unsupervised and conditional cases [2, 33, 35, 36]. Figure 1 shows results from cGAN as done in [14], where we observe some of the aforementioned issues. Cycle-GAN [31, 52] is another conditional adversarial model basing on the concept of cycle-consistency. However, such models usually require an encoding image as the condition variable, hence it is not applicable in our case.

VAEs [24] parameterize the inference and generation paths of a directed graphical model with DNNs and are trained to maximize the corresponding variational lower bounds of the data log-likelihood via stochastic backpropogation. A conditional VAE [25] introduces a new node to the graphical model as conditions to assist and constraint the generation process. Despite the many merits of this method, e.g. training stability, fast inference, GPU efficiency, etc, the combination of KL regularization and reconstruction objectives in the VLB [7, 46] often result in blurry generations—Figure 1 displays an example of attribute-conditioned face generation [48].

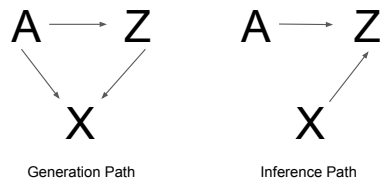


Figure 3: The generation and inference paths for the graphical model used in our work.

Many research efforts have been made to alleviate VAEs' generation blurriness. One approach is to enrich the posterior/prior distributions or the model architecture itself [9, 16, 17, 23, 41, 45]. In our work, we opt to enhance with the channel-recurrent autoencoding framework [45], motivated by its resulting latent semantics. The other approach combines VAE with autoregressive modules [18] or adversarial training [28], in the hope of taking the best from both worlds. Here we also employ an adversarial objective under a conditional setup, similar to [4]. But [4] generates based on one-hot identity label whereas we tackle with the high-dimensional attribute information.

3. Preliminaries

In this section, we first elaborate our cVAE formulation along with its graphical model, followed by an introduction of cGAN as well as their combination cVAE-GAN.

3.1. conditional VAE and Graphical Model

An important building block in our work is the conditional variational autoencoder (cVAE) [25, 48]. In our work, we choose to follow the graphical model described in Figure 3, where the attributes A are always given in our case, Z are the latent variables and X images. The intuition behind our graphical model is to construct meaningful submanifolds within the latent space, where each submanifold associates with a designated combination of attributes. In other words, our graphical model choice assumes that $p(z|a)$ is more feasible than $p(z)$ as latent distributions, where a is the attribute condition. The variational lower bound can be derived as

$$
\begin{aligned}
&\log(p(x|a)) = \log \int p(x,z|a)dz \\
&= \log \int p(x|a,z)p(z|a)\frac{q(z|x,a)}{q(z|x,a)}dz \\
&= \log \int p(x|a,z)\frac{p(z|a)}{q(z|x,a)}q(z|x,a)dz \\
&\geq \mathbb{E}_{q(z|x,a)}(\log p(x|a,z) - \log \frac{q(z|x,a)}{p(z|a)}) \\
&= -KL(q(z|x)||p(z|a)) + \mathbb{E}_{q(z|x,a)}[\log p(x|a,z)] = \mathcal{L}_{\text{cVAE}},
\end{aligned}
$$

where the approximate posterior $q(z|x)$ and the prior $p(z|a)$ are modeled as diagonal Gaussians. As done in [48], we model the latent space as 1-dim vector space $z \in \mathbb{R}^c$, which is connected via fully-connected (FC) layers. The prior network also connects attribute vectors—an N-dim vector composed of ± 1 entries where N is the number of attributes, —to the distribution of $p(z|a)$ via an FC layer. The generations of cVAEs, as shown in Figure 1 and [48], are very blurry.

3.2. conditional GAN

In [14], the author uses attribute-conditioned cGAN to generate faces: on top of GAN, cGAN feeds attribute information $a(x)$ to the generator G and discriminator D. Also, the discriminator not only needs to detect generated images but also mismatched image-attribute pairs, which yields the following over-all game played by D and G:

$$\begin{aligned}\min_G \max_D V(D,G) =& \mathbb{E}_{x\sim X}[\log D(x, a(x))] \\ &+\mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z,a),a))] \\ &+\mathbb{E}_{x,x'\sim X, x\neq x'}[\log(1 - D(x, a(x')))],\end{aligned}$$

where z is sampled from a standard Gaussian distribution, a is the attribute condition and $a(x)$ is the attribute for x. The output images indeed reflect the specified attributes but they contain many artifacts (Figure 1). Also, cGAN can suffer from low probability region coverage.

3.3. conditional VAE-GAN

The combination of VAE and GAN, VAE-GAN [28] is proposed to take the best from both worlds: high input space coverage from VAE and sharp generations from GAN. In [4], a conditional version of VAE-GAN is crafted but limited to a single-class condition, i.e. identity label. But we have multiple attribute classes associated with a single image. Therefore, we propose our version of conditional VAE-GAN (cVAE-GAN), by directly combining cVAE from Section 3.1 and cGAN from Section 3.2 to obtain the following objective:

$$\begin{aligned}\max \mathcal{L}_{\text{cVAE}} + \beta\mathbb{E}_{z\sim\{q(z|x,a(x)),p(z|a)\}}[\log D(z,a)] \\ \max_D \mathbb{E}_{x\sim X}[\log D(x,a(x))] \\ +\mathbb{E}_{z\sim\{q(z|x,a(x)),p(z|a)\}}[\log(1-D(z,a)] \\ +\mathbb{E}_{x,x'\sim X, x\neq x'}[\log(1 - D(G(x,a(x'))))], \qquad (1)\end{aligned}$$

where β is the hyperparameter to weight the gradients from the adversarial loss. The model layout is shown in Figure 2(a) and generation examples in Figure 1, where the image quality is improved yet still distant from being realistic. As cVAE and cGAN are clearly inferior to cVAE-GAN in terms of visual quality, we regard cVAE-GAN as our main baseline to compare with for the rest of the paper.

Step	Attributes
1	Heavy Makeup, Pale Skin, Rosy Cheeks
2	Brown Hair, Pointy Nose, Straight Hair
3	Arched Eyebrows, Attractive, Blond Hair
4	Blurry, Double Chin, High Cheekbones, Mouth Slightly Open, No Beard, Bags under Eyes
5	5'o clock shadow, Big Lips, Bushy Eyebrows, Chubby, Goatee, Gray Hair, Oval Face, Necktie
6	Bangs, Black Hair, Mustache, Receding Hairline, Smiling, Wavy Hair, Earrings, Hat
7	Bald, Eyeglasses, Male, Narrow Eyes
8	Big Nose, Lipstick

Table 1: In acVAE-GAN, each attribute attends a specific block.

4. Method

This section introduces the conditional channel-recurrent VAE-GAN (crVAE-GAN) and proposes an attention module such that each latent block focuses on a subset of attributes, arriving at the attentive conditional crVAE-GAN (acVAE-GAN), followed by description of how to increase generation resolution via progressive-growth training.

4.1. Channel-Recurrency

To enhance the latent space construction which in turn improves generation quality, [45] proposes the channel-recurrent architecture with which the inference and generation paths connect to the latent space. The distributions for the approximate posteriors and priors are still diagonal Gaussians, but since channel-recurrency imposes more complex manipulation to the latent space, the features, captured in a global-to-local, coarse-to-fine manner, become more abstract with more interpretable semantics.

Here, we integrate this technique to cVAE-GAN to similarly improve the conditional synthesis. The latent space is now a 3-dim space $z \in \mathbb{R}^{c\times w\times h}$. The prior network still connects the attribute vectors to the prior latent space via FC layers and reshape the resulting distributions to $\mu^p, \sigma^p|a \in \mathbb{R}^{c\times w\times h}$ (Figure 2(b)). Before being sent to the LSTM module, an attribute vector is simply repeated T times, i.e. $[a_1, a_2, \cdots a_T], a{=}a_i$, where T is the number of time steps. During generation, $z{=}[z_1, \cdot\cdot, z_T]$, with $z_i{\in}\mathbb{R}^{w\times h\times \frac{c}{T}}$, are sampled from $\mathcal{N}(\mu^p, \sigma^p)$ and concatenated with a_i at each time step i before passing through an LSTM to obtain a transformed representation $u = [u_1, \cdot\cdot, u_T]{=}\text{LSTM}(z, a)$, which is then projected back to the pixel space by a decoder. Similarly, during inference, we first transform the attribute vector with an FC layer $a'{=}\text{FC(a)} \in \mathbb{R}^w{\times}h{\times}\frac{c}{T}$, repeat the outputs T times, $[a'_1, a'_2, \cdots a'_T]$, and concatenate each to the T corresponding blocks within the convolutional features from the encoder. Then, the mean path performs convolution on the concatenated features; the variance path slices features back into T blocks, each referred to as σ_i, and feeds σ_i's into another LSTM to output σ_i^{rnn}.

4.2. Attending Attributes

Repeating the same attribute vector T times for all of the LSTM time steps can appear redundant. To obtain better interpretable features, it is desirable to understand that for each block, which attributes are being modulated, motivating us to propose an *Attention Operator* over the attribute

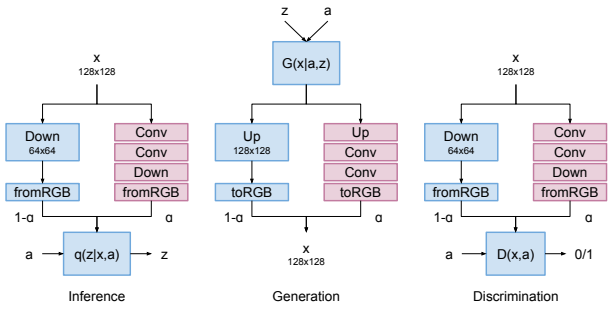


Figure 4: Illustration of progressive growing of acVAE-GAN. 64×64 acVAE-GAN is first trained without up/down sampling layers and used to initialize blue boxes of PG acVAE-GAN. The model is trained to generate 128×128 images with α being linearly interpolated from 0 to 1 over the first half of the training and remains 1 for the rest.

vectors. In essence, we learn a $T\times N$ mask matrix $\mathbf{M}$. Each row vector is constrained to be an T-dim probability vector, which is additionally regularized by minimizing its entropy $\mathcal{H}(\mathbf{M}_n)$ so that ideally only one out of the T slots has high probability. In other words, each attribute is learned to primarily focuses on one of the blocks.

Concretely, the Attention Operator first repeats the attribute vector T times to form a $T\times N$ attribute matrix $\mathbf{A}$, performs element-wise multiplication $\mathbf{A}\odot\mathbf{M}$, and separates the row vectors into $[a_1, a_2, \cdots, a_T]$. Thus instead of using the same attribute vector to all LSTM steps, we input a_t to the t-th step, reflecting the attributes attending this particular time step. Ideally, the attention from the attributes is divided evenly to all time steps: it is desirable for each time step to take in approximately N/T attributes rather than having a single time step getting most of the attribute information. Luckily, we discover no additional regularization is required to achieve such effects. For example, we summarize the attribute attention from our 8-time step acVAE-GAN in Table 1. The prior network (Figure 2(b)) now can also be meaningfully modeled with an LSTM where the tth time step takes a_t and outputs μ_i^p, σ_i^p. The remaining of acVAE-GAN are the same as crVAE-GAN.

4.3. Progressive Growing Conditional Synthesis

While our model can generate 64×64 face images from attributes, generating higher-resolution still remains a challenge. Following the idea of stackGAN [51], [45] achieves higher-resolution outputs by employing an upsampling network that takes generated low-resolution images as input. One drawback of such approach, however, is that the performance of the upsampling network is significantly limited by the initial image generation module as it is trained greedily without finetuning the entire system. Furthermore, the upsampling network usually is again composed of multiple convolution and deconvolution layers, making the whole pipeline computationally inefficient.

Instead, we borrow the idea of progressive growing GAN (PGGAN) [21] for attribute-conditioned high-resolution image generation. Specifically, we first train a model to generate 64×64 images. Then, for the generation network, we append a nearest-neighbor upsampling layer and two 3×3 convolution followed by one image decoding layer (1×1 convolution with 3 output channels) to the last feature response map to generate 128×128 images. The original and generated images of size 128×128 are then fed to the encoder and discriminator heads, respectively, which are composed of 1×1 convolution followed by two 3×3 convolutions and average pooling layer with pool size of 2 to generate 64×64 feature response maps. We illustrate the construction of progressive growing acVAE-GAN in Figure 4.

Similarly to [21], we train our progressive growing model by linearly interpolating the generated images of naive upsampling and the upsampling network. Note that our approach does not require to start from extremely low resolution images (e.g., 4×4) as 64×64 generations can be directly achieved with decent quality. The overall training objective still follows that in Equation (1).

5. Experiments

We introduce the dataset used in our experiments and describe important implementation details. For evaluation, we generate 64×64 attribute-conditioned face images, then progressively grow the resolution to 128×128. We perform qualitative visual examination and quantitative assessments namely the inception scores, human preferences, and the attribute classification accuracy. Finally, we conduct more qualitative analysis such as conditional latent space interpolation and progressively adding attributes to an image.

5.1. CelebA Dataset and Implementation Details

Our experiments are on the CelebA [30] dataset, containing $163,770$ training, $19,867$ validation, and $19,962$ testing images of face. The face ROIs are cropped and scaled to $64\times64/128\times128$. Random horizontal flipping is used during training as data augmentation. There are 40 binary attributes annotated in CelebA (see Table 1), making it a perfect venue for our experiments.

There have been many works to improve the stability of adversarial training, many of which we have experimented with, such as projective discriminator [36], self-attention [50], hinge adversarial loss [29], etc. In the end, our model adapts three main methods to stablize the optimization: batch discriminator [42], controlled discriminator update [27], and mutual information regularization [45].

The inference, generation and discriminator networks of our models share a common convolutional encoder/decoder architecture, composed of convolutional layers, batch normalization layers [19], and activation layers [32, 44]. See the code in the Supplemental Materials for details.

Stocahstic gradinet descent is done using ADAM [22] with $\epsilon = 1\times10^{-8}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ for 150 epochs.

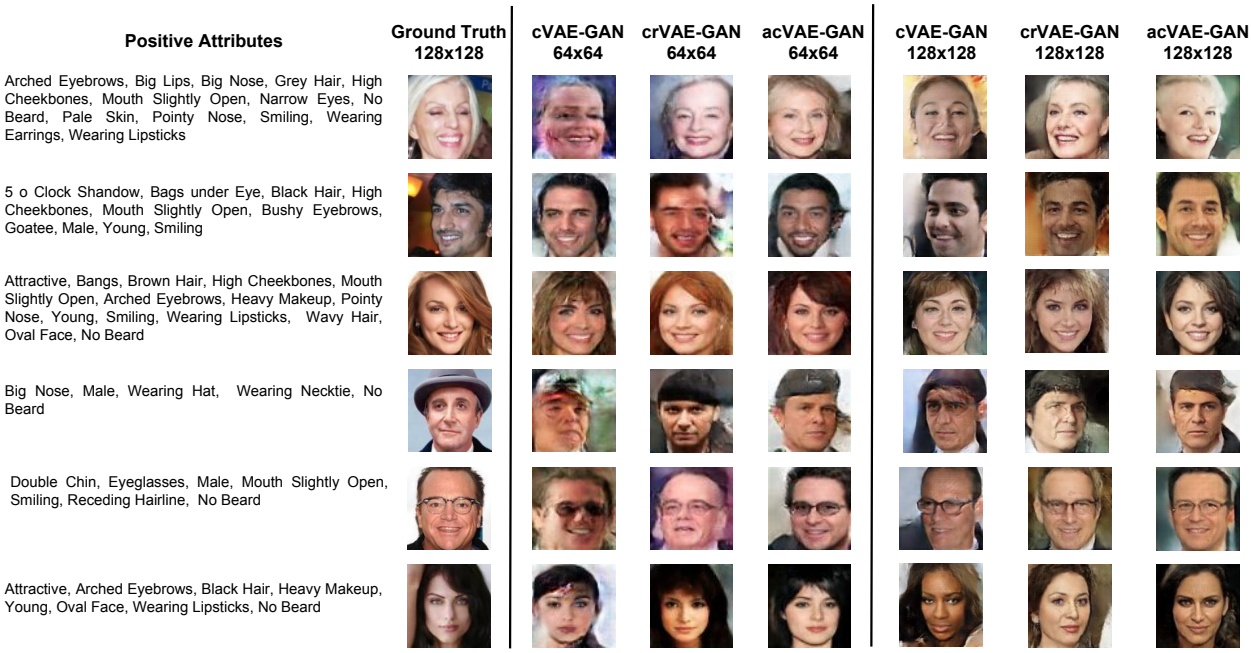


Figure 5: We select attribute combinations from the testing set and compare 64×64 and 128×128 resolution image generations of cVAE-GAN (baseline), crVAE-GAN and acVAE-GAN, along with the ground truth images.

Models	64×64	128×128	128×128
cVAE-GAN	37.48±0.96	93.74±2.25	23%
crVAE-GAN	54.66±0.67	87.51±1.46	28%
acVAE-GAN	**55.12**±0.64	**102.23**±1.81	**47**%

Table 2: Inception Scores on 64×64/128×128 generated images from testing set attributes. We perform 10-fold calculation and report the mean±std. We also report the percentage humans prefer a model in a 3-way classification setup.

The initial learning rate is 0.0003 for (cVAE-GAN) and 0.001 for the rest. We follow the control protocol in [27] to update the discriminator with a variation of ADAM that imposes a threshold for updating: if the classification accuracy for a batch consisting of a third real pairs, a third fake pairs and a third mis-matched pairs is over 90%, we skip the update. During progressive-growth training, the resolution transition is linearly done in 75 epochs. Training code is in the Supplemental Materials.

5.2. Attribute-Conditioned Face Synthesis

Evaluation of the conditional generative models in our work primarily focuses on the visual fidelity and the faithfulness to the assigned attributes. To this end, we first visually compare the baseline cVAE-GAN with crVAE-GAN and acVAE-GAN in Figure 5, where our proposed model produces more photo-realistic generations that satisfy the attribute conditions, even for some of the more challenging ones such as eyeglasses and hats.

We also measure the inception scores [42] for 64×64 and 128×128 images. The classification model used here is a VGG11 trained on CASIA [49] for face recognition with input resolution 128×128. During testing, we generate images conditioning on all attribute combinations from the testing set, randomly divide them to 10 subsets, obtain inception scores for these subsets and report their mean±standard deviation in Table 2. However, since inception score is not an absolutely truthful metric on visual quality [5], we present 200 128×128 generated tuples from the 3 models with the same set of attributes to mechanical turkers, out of which they select the one with the highest visual quality. The results are also summarized in Table 2. Our proposed model substantially outperform the baseline in both metrics.

To quantitatively judge the attribute qualities, we train an attribute classifiers on 64×64/128×128 CelebA images. Our classifier is composed of a standard convolutional encoder and a fully connected layer ending with 40 dimension binary predictions for the 40 attributes. We follow the standard training-validation procedures and report the testing results in Table 3. Note that our classifier achieves competitive accuracy comparing [30]. Next we run our attribute classifiers on the generated images, with accuracy results in Table 3. All of the models reach promising accuracy, meaning that our graphical model design is very effective in delivering the attributes, among which our proposed model has the highest score.

5.3. Content Interpolation

For our goal of generating diverse and realistic images given a set of attributes, it is important to learn a prior latent space that densely covers valid images in the pixel space conditioned on the attributes. In this section, we investigate this quality of the latent spaces learned by our models

		5 Shadow	Arch. Eyebrows	Attractive	Bags Un. Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	H. Cheekbones	Male
64×64	Real Images	93	81	81	84	98	95	71	82	88	95	95	88	92	95	96	99	96	98	90	87	97
	cVAE-GAN	92	78	80	82	98	96	69	79	94	95	94	82	91	94	95	98	96	97	91	86	98
	crVAE-GAN	93	79	81	83	98	97	70	83	91	96	94	85	93	95	95	99	96	98	91	88	98
	acVAE-GAN	92	81	81	84	98	97	71	83	87	96	94	85	93	95	95	99	96	98	90	88	99
128×128	Real Images	94	83	82	84	98	96	70	84	89	95	95	88	92	95	96	99	97	98	91	87	97
	cVAE-GAN	92	83	84	83	98	98	70	83	87	97	95	87	93	95	95	99	96	98	92	88	99
	crVAE-GAN	91	81	81	83	98	98	70	83	89	97	94	85	93	95	95	99	96	98	91	88	98
	acVAE-GAN	92	83	84	83	98	98	70	83	87	97	95	87	93	95	95	99	96	98	92	88	99

		Mouth S. O.	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose	Reced. Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wear. Earrings	Wear. Hat	Wear. Lipstick	Wear. Necklace	Wear. Necktie	Young	Average
64×64	Real Images	92	96	86	94	74	96	76	93	95	97	92	81	80	87	98	93	86	94	87	90.4
	cVAE-GAN	94	95	86	95	70	95	72	93	90	96	94	77	69	81	98	94	86	94	84	89.1
	crVAE-GAN	96	96	87	95	73	96	73	93	94	96	95	78	78	84	98	94	86	94	88	90.3
	acVAE-GAN	97	96	89	95	72	96	74	93	94	96	96	79	76	84	98	94	86	94	86	**90.4**
128×128	Real Images	93	96	87	95	75	97	77	93	94	97	92	81	82	89	99	94	86	95	88	90.9
	cVAE-GAN	98	96	89	96	75	96	76	93	93	96	97	80	79	88	99	94	85	95	87	90.7
	crVAE-GAN	97	96	88	94	73	96	74	93	93	96	96	79	77	87	98	94	84	94	87	90.6
	acVAE-GAN	98	96	89	96	75	96	76	93	93	96	97	80	79	88	99	94	85	95	87	**91.0**

Table 3: We train an independently trained classifier, we test attribute classification results on real images as benchmark and generated ones from the baseline cVAE-GAN, crVAE-GAN and acVAE-GAN. All models perform competatively, likely due to the adversarial loss and the proposed graphical model. All numbers are in %.

Figure 6: We generate faces by linearly interpolating latent variable between 2 samples drawn from the conditional priors conditioned with the same attributes. The conditional prior latent space in all models appear to deliver smooth transitioning while maintaining the attributes; acVAE-GAN especially excels at traversing across diverse samples and at the same time maintaining image qualities.

by checking how well two distinct modes on the latent submanifold conditioned on the same attributes can mix, also an indicator of the representation abstractness [6]. Concretely, we select four sets of attribute combinations from the testing set, obtain the conditional prior space for these combinations, and sample 2 examples each to represent 2 modes Z_1 and Z_2 on the latent submanifold. Then we traverse from Z_1 to Z_2 using the interpolation formula

$$Z_i = \cos^2(\psi)Z_1 + (1 - \cos^2(\psi))Z_2, \psi \in [0, \pi/2].$$

The generation network projects the interpolated latent codes to the pixel space. Figure 6 shows, likely thanks to our proposed graphical model learning conditional submanifold distributions, the resulting latent representations from all models can be consistently mixed in a semantically meaningful manner. Our proposed acVAE-GAN especially excels at transiting through very distinct looking modes, such as the pose variations in the 2nd row, the glasses changing from clear to dark in the 3rd row and the background color shifting from the last row. Meanwhile, some of the generations from the other models associated with certain attributes appear to be repetitive, such as the 2nd row of cVAE-GAN and crVAE-GAN; some other generations from cVAE-GAN are prone to artifacts such as the 3rd and 4th row.

5.4. Generation Progression

Our proposed acVAE-GAN has each attribute to focus its attention to a specific time-step latent block to achieve more interpretable high-level features, in the sense that we in theory can pinpoint the exact block modulating a given attribute. To verify this assertion, we at-

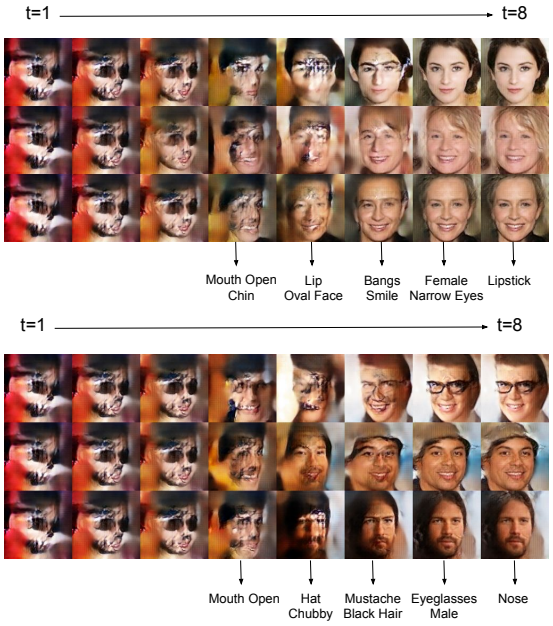


Figure 7: We progressively construct the conditional prior latent space by sending in the attended attributes for each time step, and then fill in the latent representations sequentially. The first few time steps do not carry enough content information to distinctively show the progression; for the latter ones, after sampling each latent block, its associated attributes indeed appear (correspondence see Table 1). We list out a few examples such as gender, hair color, etc.

tempt to visualize the emergence of attributes by progressively sampling each latent blocks. Recall that the conditional prior space is generated via sending the attended attribute a_t through an LSTM module at tth time-step, which then outputs the distribution $\mathcal{N}(\mu_t^p, \sigma_t^p)$ of the corresponding prior latent block. From $\mathcal{N}(\mu_t^p, \sigma_t^p)$, we sample z_t and, in combination of the previously sampled z's, decode $[(z_1, a_1), \cdots (z_i, a_i), (0, 0), \cdots, (0, 0)]$ via the generation network back to the pixel space. We plot the $t = 1 \cdots 8$ generations in Figure 7. The first 3 time steps do not carry enough content information to distinctively show the progression. However, for the latter ones, after sampling each z_t, its associated attributes a_t do emerge (attribute correspondence see Table 1). For example, we can already see mouth open in step 3; hat forms in step 4 and glasses step 7; the gender is finalized at step 7; mustache and bangs become clear at step 6; step 8 brings subtle changes over the lip color and nose shape. The visualization demonstrates that up to a certain degree we can confirm a latent block is indeed modulating its attending attributes.

6. Application

The recurrent learning of latent blocks enable each of them to be in charge of a different aspect of the input content [45]. In particular, some latent blocks can impose more drastic global impact and some others in a more subtle way. We leverage this unique feature from channel-recurrency and propose the following 2-stage generation pipeline. As

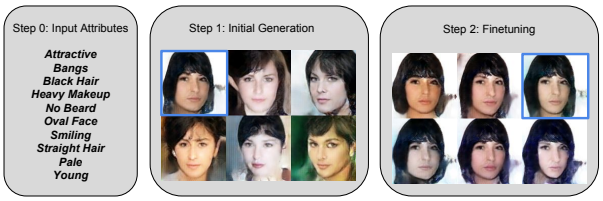


Figure 8: A 2-stage generation tool where the model first outputs a diversity of faces following the given attributes and then, after selecting an initial output, resamples one of the blocks in charge of minor modifications to give more finetuned options.

a preparation step, through trial-and-error, we locate the latent block that perform minor adjustment on the content of the generation; in the case of the model used in Figure 7, a progressive-growing acVAE-GAN, is z_5. The first step generates a diversity of samples from the corresponding prior latent space based on given attributes and the user selects the most desirable one. For finetuining, the 2nd step resamples multiple z_5 of the selected example while fixing the rest of z_t's and the user finalizes on the most closely-matching image. An example is demonstrated in Figure 8 where the first step outputs distinct faces obeying the same set of attributes and the second step finetunes face shape, skin tone, hair details, etc. The user can further use existing neural editing tools [8] to refine the image, but it is out of scope of this work.

7. Conclusion

We propose an attentive conditional channel-recurrent VAE-GAN for high-quality attribute-to-face synthesis while learning better interpretable high-level features. We also incorporate progressive-growth training to generate higher-resolution images. We demonstrate the superiority of our models both in quantitative and qualitative evaluations. In application, we envision a tool for a general-to-specific 2-stage attribute-conditioned face synthesis. Future research includes extending our framework in a semi-supervised manner and extrapolating our models to other tasks.

Acknowledgments

We are grateful to Herke van Hoof and Max Welling for their valuable comments and NVIDIA for the donation of GPUs.

References

[1] Faces 4.0. http://www.iqbiometrix.com, 2016. 1
[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv*, 2017. 3
[3] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (gans). *ICML*, 2017. 3
[4] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Cvae-gan: fine-grained image generation through asymmetric training. 2, 3, 4

[5] S. Barratt and R. Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018. 6

[6] Y. Bengio, G. Mesnil, Y. Dauphin, and S. Rifai. Better mixing via deep representations. In *ICML*, 2013. 7

[7] O. Bousquet, S. Gelly, I. Tolstikhin, C.-J. Simon-Gabriel, and B. Schoelkopf. From optimal transport to generative modeling: the vegan cookbook. *arXiv*, 2017. 3

[8] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016. 8

[9] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. *arXiv*, 2016. 3

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, June 2018. 3

[11] FBI. Preliminary semiannual uniform crime report, januaryjune, 2017, 2017. 1

[12] C. D. Frowd, D. Carson, H. Ness, D. McQuiston-Surrett, J. Richardson, H. Baldwin, and P. Hancock. Contemporary composite techniques: The impact of a forensically-relevant target delay. *Legal and Criminological Psychology*, 10(1):63–81, 2005. 1

[13] H. Gao, H. Yuan, Z. Wang, and S. Ji. Pixel deconvolutional networks. *arXiv preprint arXiv:1705.06820*, 2017. 3

[14] J. Gauthier. Conditional generative adversarial nets for convolutional face generation. 1, 3, 4

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 3

[16] K. Gregor, F. Besse, D. J. Rezende, I. Danihelka, and D. Wierstra. Towards conceptual compression. In *NIPS*, 2016. 3

[17] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, 2015. 3

[18] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville. Pixelvae: A latent variable model for natural images. *arXiv*, 2016. 3

[19] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[20] T. Kaneko, K. Hiramatsu, and K. Kashino. Generative attribute controller with conditional filtered generative adversarial networks. In *CVPR*, July 2017. 3

[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability and variation. *arXiv*, 2017. 2, 5

[22] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[23] D. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improving variational inference with inverse autoregressive flow. In *NIPS*, 2016. 3

[24] D. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2013. 1, 3

[25] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *NIPS*, 2014. 1, 3

[26] S. Klum, H. Han, A. K. Jain, and B. Klare. Sketch based face recognition: Forensic vs. composite sketches. In *International Conference on Biometrics (ICB)*, 2013. 1

[27] A. B. L. Larsen and S. K. Sønderby. Generating faces with torch, 2016. 5, 6

[28] A. B. L. Larsen, S. K. Sonderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016. 2, 3, 4

[29] J. H. Lim and J. C. Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 5

[30] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5, 6

[31] Y. Lu, Y.-W. Tai, and C.-K. Tang. Conditional cyclegan for attribute guided face image generation. *arXiv preprint arXiv:1705.09966*, 2017. 3

[32] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop*, 2013. 5

[33] X. Mao, Q. Li, H. Xie, R. Y. Lau, and Z. Wang. Multi-class generative adversarial networks with the l2 loss function. *arXiv*, 2016. 3

[34] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[35] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 3

[36] T. Miyato and M. Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 3, 5

[37] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *ICML*, 2016. 3

[38] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional image generation with pixelcnn decoders. *NIPS*, 2016. 1, 3

[39] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *ICML*, 2016. 3

[40] S. Reed, A. v. d. Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, D. Belov, and N. de Freitas. Parallel multiscale autoregressive density estimation. *arXiv preprint arXiv:1703.03664*, 2017. 1, 3

[41] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *ICML*, 2015. 3

[42] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016. 5, 6

[43] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv*, 2017. 3

[44] W. Shang, K. Sohn, D. Almeida, and H. Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *ICML*, 2016. 5

[45] W. Shang, K. Sohn, and Y. Tian. Channel-recurrent autoencoding for image modeling. *WACV*, 2018. 2, 3, 4, 5, 8

[46] L. Theis, A. v. d. Oord, and M. Bethge. A note on the evaluation of generative models. In *ICLR*, 2016. 3

[47] Y. Wu, Y. Burda, R. Salakhutdinov, and R. Grosse. On the quantitative analysis of decoder-based generative models. *arXiv*, 2016. 3

[48] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, 2016. 1, 3, 4

[49] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv*, 2014. 6

[50] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-

attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. 5

[51] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv*, 2016. 5

[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017. 3