Robust Conditional GAN from Uncertainty-Aware Pairwise Comparisons

Ligong Han,[1] **Ruijiang Gao,**[2] **Mun Kim,**[1] **Xin Tao,**[3] **Bo Liu,**[4] **Dimitris Metaxas**[1]

[1]Department of Computer Science, Rutgers University
[2]McCombs School of Business, The University of Texas at Austin
[3]Tencent YouTu Lab, [4]JD Finance America Corporation
{l.han, mun.kim}@rutgers.edu, ruijiang@utexas.edu
{jiangsutx, kfliubo}@gmail.com, dnm@cs.rutgers.edu

Abstract

Conditional generative adversarial networks have shown exceptional generation performance over the past few years. However, they require large numbers of annotations. To address this problem, we propose a novel generative adversarial network utilizing weak supervision in the form of pairwise comparisons (PC-GAN) for image attribute editing. In the light of Bayesian uncertainty estimation and noise-tolerant adversarial training, PC-GAN can estimate attribute rating efficiently and demonstrate robust performance in noise resistance. Through extensive experiments, we show both qualitatively and quantitatively that PC-GAN performs comparably with fully-supervised methods and outperforms unsupervised baselines. Code and Supplementary can be found on the project website[*].

Introduction

Generative adversarial networks (GAN) (Goodfellow et al. 2014) have shown great success in producing high-quality realistic imagery by training a set of networks to generate images of a target distribution via an adversarial setting between a generator and a discriminator. New architectures have also been developed for adversarial learning such as conditional GAN (CGAN) (Mirza and Osindero 2014; Odena, Olah, and Shlens 2016; Han, Murphy, and Ramanan 2018) which feeds a class or an attribute label for a model to learn to generate images conditioned on that label. The superior performance of CGAN makes it favorable for many problems in artificial intelligence (AI) such as image attribute editing.

However, this task faces a major challenge from the lack of massive labeled images with varying attributes. Many recent works attempt to alleviate such problems using semi-supervised or unsupervised conditional image synthesis (Lucic et al. 2019). These methods mainly focus on conditioning the model on categorical pseudo-labels using self-supervised image feature clustering. However, attributes are often continuous-valued, for example, the stroke thickness of MNIST digits. In such cases, applying unsupervised clustering would be difficult since features are most likely to be grouped by salient attributes (like identities) rather than
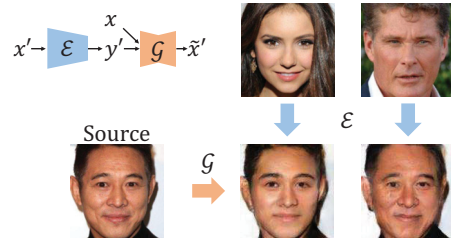
[*]https://github.com/phymhan/pc-gan

Figure 1: The generative process. Starting from a source image x, our model is able to synthesize a new image $\tilde{x}'$ with the desired attribute intensity possessed by the target image x'.

any other attributes of interest. In this work, to disentangle the target attribute from the rest, we focus on learning from weak supervisions in the form of pairwise comparisons.

Pairwise comparisons. Collecting human preferences on pairs of alternatives, rather than evaluating absolute individual intensities, is intuitively appealing, and more importantly, supported by evidence from cognitive psychology (Fürnkranz and Hüllermeier 2010). As pointed out by Yan (2016), we consider relative attribute annotation because they are (1) easier to obtain than total orders, (2) more accurate than absolute attribute intensities, and (3) more reliable in application like crowd-sourcing. For example, it would be hard for an annotator to accurately quantify the attractiveness of a person's look, but much easier to decide which one is preferred given two candidates. Moreover, attributes in images are often subjective. Different annotators have different criteria in their mind, which leads to noisy annotations (Xu et al. 2019).

Thus, instead of assigning an absolute attribute value to an image, we allow the model to learn to rank and assign a relative order between two images (Yan 2016; Fürnkranz and Hüllermeier 2010). This method alleviates the aforementioned problem of lacking continuously valued annotations by learning to rank using pairwise comparisons.

Weakly supervised GANs. Our main idea is to substitute the full supervision with the attribute ratings learned from weak supervisions, as illustrated in Figure 1. To do so, we draw inspiration from the Elo rating system (Elo 1978) and

design a Bayesian Siamese network to learn a rating function with uncertainty estimations. Then, for image synthesis, motivated by (Thekumparampil et al. 2018) we use "corrupted" labels for adversarial training. The proposed framework can (1) learn from pairwise comparisons, (2) estimate the uncertainty of predicted attribute ratings, and (3) offer quantitative controls in the presence of a small portion of absolute annotations. Our contributions can be summarized as follows.

- We propose a weakly supervised generative adversarial network, PC-GAN, from pairwise comparisons for image attribute manipulation. To the best of our knowledge, this is the first GAN framework considering relative attribute orders.

- We use a novel attribute rating network motivated from the Elo rating system, which models the latent score underlying each item and tracks the uncertainty of the predicted ratings.

- We extend the robust conditional GAN to continuous-value setting, and show that the performance can be boosted by incorporating the predicted uncertainties from the rating network.

- We analyze the sample complexity which shows that this weakly supervised approach can save annotation effort. Experimental results show that PC-GAN is competitive with fully-supervised models, while surpassing unsupervised methods by a large margin.

Related Work

Learning to rank. Our work focuses on finding "scores" for each item (e.g. player's rating) in addition to obtaining a ranking. The popular Bradley-Terry-Luce (BTL) model postulates a set of latent scores underlying all items, and the Elo system corresponds to the logistic variant of the BTL model. Numerous algorithms have been proposed since then. To name a few, TrueSkill (Herbrich, Minka, and Graepel 2007) considers a generalized Elo system in the Bayesian view. Rank Centrality (Negahban, Oh, and Shah 2016) builds on spectral ranking and interprets the scores as the stationary probability under the random walk over comparison graphs. However, these methods are not designed for amortized inference, i.e. the model should be able to score (or extrapolate) an unseen item for which no comparisons are given. Apart from TrueSkill and Rank Centrality, the most relevant work is the RankNet (Burges et al. 2005). Despite being amortized, RankNet is homoscedastic and falls short of a principled justification as well as providing uncertainty estimations.

Weakly supervised learning. Weakly-supervised learning focuses on learning from coarse annotations. It is useful because acquiring annotations can be very costly. A close weakly supervised setting to our problem is (Xiao and Jae Lee 2015) which learns the spatial extent of relative attributes using pairwise comparisons and gives an attribute intensity estimation. However, most facial attributes like attractiveness and age are not localized features thus cannot be exploited by local regions. In contrast, our work uses this relative attribute intensity for attribute transfer and manipulation.

Uncertainty. There are two uncertainty measures one can model: aleatoric uncertainty and epistemic uncertainty. The epistemic uncertainty captures the variance of model predictions caused by lack of sufficient data; the aleatoric uncertainty represents the inherent noise underlying the data (Kendall and Gal 2017). In this work, we leverage Bayesian neural networks (Gal and Ghahramani 2016) as a powerful tool to model uncertainties in the Elo rating network.

Robust conditional GAN (RCGAN). Conditioning on the estimated ratings, a normal conditional generative model can be vulnerable under bad estimations. To this end, recent research introduces noise robustness to GANs. Bora, Price, and Dimakis (2018) apply a differentiable corruption to the output of the generator before feeding it into the discriminator. Similarly, RCGAN (Thekumparampil et al. 2018) proposes to corrupt the categorical label for conditional GANs and provides theoretical guarantees. Both methods have shown great denoising performance when noisy observations are present. To address our problem, we extend RCGAN to the continuous-value setting and incorporate uncertainties to guide the image generation.

Image attribute editing. There are many recent GAN-style architectures focusing on image attribute editing. IPC-GAN (Wang et al. 2018) proposes an identity preserving loss for facial attribute editing. Zhu et al. (2017) propose cycle consistency loss that can learn the unpaired translation between image and attribute. BiGAN/ALI (Donahue, Krähenbühl, and Darrell 2016; Dumoulin et al. 2016) learns an inverse mapping between image-and-attribute pairs.

There exists another line of research that is not GAN-based. Deep feature interpolation (DFI) (Upchurch et al. 2017) relies on linear interpolation of deep convolutional features. It is also weakly-supervised in the sense that it requires two domains of images (e.g. young or old) with inexact annotations (Zhou 2017). DFI demonstrates high-fidelity results on facial style transfer. While, the generated pixels look unnatural when the desired attribute intensity takes extreme values, we also find that DFI cannot control the attribute intensity quantitatively. Unlike prior research, our method uses weak supervision in the form of pairwise comparisons and leverages uncertainty together with noise-tolerant adversarial learning to yield a robust performance in image attribute editing.

Pairwise Comparison GAN

In this section, we introduce the proposed method for pairwise weakly-supervised visual attribute editing. Denote an image collection as $I = \{x_1, \cdots, x_n\}$ and x_i's underlying absolute attribute values as $\Omega(x_i)$. Given a set of pairwise comparisons C (e.g., $\Omega(x_i) > \Omega(x_j)$ or $\Omega(x_i) = \Omega(x_j)$, where $i, j \in \{1, \cdots, n\}$), our goal is to generate a realistic image quantitatively with a different desired attribute intensity, for example, from 20 years old to 50 years old. The proposed framework consists of an Elo rating network followed by a noise-robust conditional GAN.
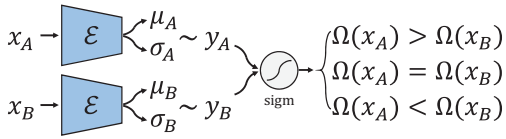
Figure 2: The Elo rating network. The comparison is performed by feeding into a sigmoid function the difference of ratings (scalar) of a given image pair. After training, the encoder $\mathcal{E}$ is used to train the PC-GAN, as illustrated in Figure 3.

Attribute Rating Network

The designed attribute rating module is motivated by the Elo rating system (Elo 1978), which is widely used to evaluate the relative levels of skills between players in zero-sum games. Elo rating from a player is represented as a scalar value which is adjusted based on the outcome of games. We apply this idea to image attribute editing by considering each image as a player and comparison pairs as games with outcomes. Then we learn a rating function.

Elo rating system. The Elo system assumes the performance of each player is normally distributed. For example, if `Player A` has a rating of y_A and `Player B` has a rating of y_B, the probability of `Player A` winning the game against `Player B` can be predicted by $P_A = \frac{1}{1+10^{(y_B-y_A)/400}}$. We use S_A to denote the actual score that `Player A` obtains after the game, which can be valued as $S_A(\texttt{win}) = 1, S_A(\texttt{tie}) = 0.5, S_A(\texttt{lose}) = 0$. After each game, the player's rating is updated according to the difference between the prediction P_A and the actual score S_A by $y'_A = y_A + K(S_A - P_A)$, where K is a constant.

Image pair rating prediction network. Given an image pair (x_A, x_B) and a certain attribute Ω, we propose to use a neural network for predicting the relative attribute relationship between $\Omega(x_A)$ and $\Omega(x_B)$. This design allows amortized inference, that is, the rating prediction network can provide ratings for both seen and unseen data. The model structure is illustrated in Figure 2.

The network contains two input branches fed with x_A and x_B. For each image x, we propose to learn its rating value y_x by an encoder network $\mathcal{E}(x)$. Assuming the rating value of x follows a normal distribution, that is $y_x \sim \mathcal{N}\left(\mu(x), \sigma^2(x)\right)$, we employ the reparameterization trick (Kingma and Welling 2013), $y_x = \mu(x) + \epsilon\sigma(x)$ (where $\epsilon \sim \mathcal{N}(0, I)$). After obtaining each image's latent rating y_A and y_B, we formulate the pair-wise attribute comparison prediction as $P_{A,y}(\Omega(x_A) > \Omega(x_B)|x_A, x_B, y_A, y_B) = \text{sigm}(y_A - y_B)$ where sigm is the sigmoid function. Then, the predictive probability of x_A winning x_B is obtained by integrating out the latent variables y_A and y_B,

$$P_A(\Omega(x_A) > \Omega(x_B)|x_A, x_B) = \int \text{sigm}(y_A - y_B)dy_A dy_B,$$
(1)

and $P_B = 1 - P_A$. The above integration is intractable, and can be approximated by Monte Carlo, $P_A \approx P_A^{MC} = \frac{1}{M}\sum_{m=1}^M P_{A,y}$. We denote the ground-truth of P_A and P_B

as S_A and S_B. The ranking loss $\mathcal{L}_{rank}$ can be formulated with a logistic-type function, that is

$$\mathcal{L}_{rank}^{MC} = -\mathbb{E}_{x_A, x_B \sim C}\left[S_A \log P_A^{MC} + S_B \log P_B^{MC}\right].$$
(2)

Noticing that $\mathcal{L}_{rank}^{MC}$ is biased, an alternative unbiased upper bound can be derived as

$$\mathcal{L}_{rank}^{UB} = -\mathbb{E}_{x_A, x_B \sim C}\left[\frac{1}{M}\sum_{m=1}^M S_A \log P_{A,y} + S_B \log P_{B,y}\right].$$
(3)

In practice, we find that $\mathcal{L}_{rank}^{UB}$ performs slightly better than $\mathcal{L}_{rank}^{MC}$.

We further consider a Bayesian variant of $\mathcal{E}$. The Bayesian neural network is shown to be able to provide the epistemic uncertainty of the model by estimating the posterior over network weights in network parameter training (Kendall and Gal 2017) . Specifically, let $q_\theta(w)$ be an approximation of the true posterior $p(w|\text{data})$ where θ denotes the parameter of q, we measure the difference between $q_\theta(w)$ and $p(w|\text{data})$ with the KL-divergence. The overall learning objective is the negative evidence lower bound (ELBO) (Kingma and Welling 2013; Gal and Ghahramani 2016),

$$\mathcal{L}_{\mathcal{E}} = \mathcal{L}_{rank} + \underbrace{D_{KL}(q_\theta(w)\|p(w|\text{data}))}_{KL}.$$
(4)

Gal and Ghahramani (2016) propose to view dropout together with weight decay as a Bayesian approximation, where sampling from q_θ is equivalent to performing dropout and the KL term in Equation 4 becomes L_2 regularization (or weight decay) on θ.

The predictive uncertainty of rating y for image x can be approximated using:

$$\hat{\sigma}^2(y) \approx \frac{1}{T}\sum_{t=1}^T \mu_t^2 - (\frac{1}{T}\sum_{t=1}^T \mu_t)^2 + \frac{1}{T}\sum_{t=1}^T \sigma_t^2$$
(5)

with $\{\mu_t, \sigma_t\}_{t=1}^T$ a set of T sampled outputs: $\mu_t, \sigma_t = \mathcal{E}(x)$.

Transitivity. Notice that the transitivity does not hold because of the stochasticity in y. If we fix $\sigma(\cdot)$ to be zero and a non-Bayesian version is used, the Elo rating network becomes a RankNet (Burges et al. 2005), and transitivity holds. However, one can still maintain transitivity by avoiding reparameterization and modeling $P_A = \text{sigm}(\frac{\mu(x_A)-\mu(x_B)}{\sqrt{\sigma^2(x_A)+\sigma^2(x_B)}})$. In practice, we find that reparameterization works better.

Conditional GAN with Noisy Information

We construct a CGAN-based framework for image synthesis conditioned on the learned attribute rating. The overall training procedure is shown in Figure 3: given a pair of images x and x', the generator $\mathcal{G}$ is trained to transform x into $\tilde{x}' = \mathcal{G}(x, y')$, such that $\tilde{x}'$ possesses the same rating $y' = \mathcal{E}(x')$ as x'. The predicted ratings can still be noisy, thus a robust conditional GAN is considered. While RC-GAN (Thekumparampil et al. 2018) is conditioned on discrete categorical labels that are "corrupted" by a confusion
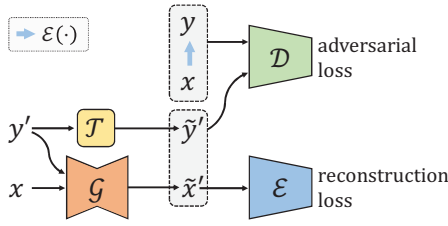
Figure 3: Overview of PC-GAN. Image $\tilde{x}'$ is synthesized from x and y'. y' is then "corrupted" to $\tilde{y}'$ by the transition $\mathcal{T}$, where $\mathcal{T}$ is a sampling process $\tilde{y}' \sim \mathcal{N}(y', \hat{\sigma}'^2)$. The reconstruction on attribute rating enforces mutual information maximization. The main difference between PC-GAN and a normal conditional GAN is that the conditioned rating of the generated sample is corrupted before feeding into the adversarial discriminator, forcing the generator to produce samples with clean ratings.

matrix, our model relies on the ratings that are continuous-valued and realizes the "corruption" via resampling.

Adversarial loss. Given image x, the corresponding rating y can be obtained from a forward pass of the pre-trained encoder $\mathcal{E}$. Thus $\mathcal{E}$ defines a joint distribution $p_{\mathcal{E}}(x, y) = p_{data}(x)p_{\mathcal{E}}(y|x)$. Importantly, the output $\tilde{x}'$ of $\mathcal{G}$ is paired with a corrupted rating $\tilde{y}' = \mathcal{T}(y')$, where $\mathcal{T}$ is a sampling process $\tilde{y}' \sim \mathcal{N}(y', \hat{\sigma}'^2)$. The adversarial loss is

$$\mathcal{L}_{CGAN} = \mathbb{E}_{x,y \sim p(x,y)} \log(\mathcal{D}(x, y)) \quad + \tag{6}$$
$$\mathbb{E}_{x \sim p(x), y' \sim p(y'), \tilde{y}' \sim \mathcal{T}(y')} \log(1 - \mathcal{D}(\mathcal{G}(x, y'), \tilde{y}')).$$

The discriminator $\mathcal{D}$ is discriminating between real data (x, y) and generated data $(\mathcal{G}(x, y'), \mathcal{T}(y'))$. At the same time, $\mathcal{G}$ is trained to fool $\mathcal{D}$ by producing images that are both realistic and consistent with the given attribute rating. As such, the Bayesian variant of the encoder is required for considering robust conditional adversarial training.

Mutual information maximization. Besides conditioning the discriminator, to further encourage the generative process to be consistent with ratings and thus learn a disentangled representation (Chen et al. 2016), we add a reconstruction loss on the predictive ratings:

$$\mathcal{L}_{rec}^y = \mathbb{E}_{x \sim p(x), y' \sim p(y')} \frac{1}{2\hat{\sigma}'^2} \|\mathcal{E}(\mathcal{G}(x, y')) - y'\|_2^2 + \frac{1}{2} \log \hat{\sigma}'^2. \tag{7}$$

The above reconstruction loss can be viewed as the conditional entropy between y' and $\mathcal{G}(x, y')$,

$$\mathcal{L}_{rec}^y \propto -\mathbb{E}_{y' \sim p(y'), \tilde{x}' \sim \mathcal{G}(x, y')} [\log p(y'|\tilde{x}')]$$
$$= -\mathbb{E}_{y' \sim p(y'), \tilde{x}' \sim \mathcal{G}(x, y')} [\mathbb{E}_{y \sim (y|\tilde{x}')} [\log (y|\tilde{x}')]]$$
$$= H(y'|\mathcal{G}(x, y')). \tag{8}$$

Following the same logic, the cycle loss can be also viewed as maximizing the mutual information between x and $\tilde{x}'$.

Full objective. Finally, the full objective can be written as:

$$\mathcal{L}(\mathcal{G}, \mathcal{D}) = \mathcal{L}_{CGAN} + \lambda_{rec}\mathcal{L}_{rec}^y + \lambda_{cyc}\mathcal{L}_{cyc}, \tag{9}$$

where λs control the relative importance of corresponding losses. The final objective formulates a minimax problem where we aim to solve:

$$\mathcal{G}^* = \arg \min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{L}(\mathcal{G}, \mathcal{D}). \tag{10}$$

Analysis of loss functions. Goodfellow et al. (2014) show that the adversarial training results in minimizing the Jensen-Shannon divergence between the true conditional and the generated conditional. Here, the approximated conditional will converge to the distribution characterized by the encoder $\mathcal{E}$. If $\mathcal{E}$ is optimal, the approximated conditional will converge to the true conditional, we defer the proof in Supplementary.

GAN training. In practice, we find that the conditional generative model trains better if equal-pairs (pairs with approximately equal attribute intensities) are filtered out and only different-pairs (pairs with clearly different intensities) are remained. Comparisons of training CGAN with or without equal-pairs can be found in Supplementary.

Number of pairs. Suppose there are n images in the dataset, then the possible number of pairs is upper bounded by $n(n-1)/2$. However, if $\mathcal{O}(n^2)$ pairs are necessary, there is no benefit of choosing pairwise comparisons over absolute label annotation. Using results from (Radinsky and Ailon 2011; Wauthier, Jordan, and Jojic 2013), the following proposition shows that only $\mathcal{O}(n)$ comparisons are needed to recover an approximate ranking. We also provide an empirical study in the Supplementary.

Proposition 0.1. *For a constant d and any $0 < \lambda < 1$, if we measure dn/λ^2 comparisons chosen uniformly with repetition, the Elo rating network will output a permutation $\hat{\pi}$ of expected risk at most $(2/\lambda)(n(n-1)/2)$.*

Experiments

In this section, we first present a motivating experiment on MNIST. Then we evaluate the PC-GAN in two parts: (1) learning attribute ratings, and (2) conditional image synthesis, both qualitatively and quantitatively.

Dataset. We evaluate PC-GAN on a variety of datasets for image attribute editing tasks:

- Annotated MNIST (Kim 2017) provides annotations of stroke thickness for MNIST (LeCun et al. 1998) dataset.

- CACD (Chen, Chen, and Hsu 2014) is a large dataset collected for cross-age face recognition, which includes 2,000 subjects and 163,446 images. It contains multiple images for each person which cover different ages.

- UTKFace (Zhang and Qi 2017) is also a large-scale face dataset with a long age span, ranging from 0 to 116 years. This dataset contains 23,709 facial images with annotations of age, gender, and ethnicity.

- SCUT-FBP (Xie et al. 2015) is specifically designed for facial beauty perception. It contains 500 Asian female portraits with attractiveness ratings (1 to 5) labeled by 75 human raters.

- CelebA (Liu et al. 2015) is a standard large-scale dataset for facial attribute editing. It consists of over 200k images, annotated with 40 binary attributes.
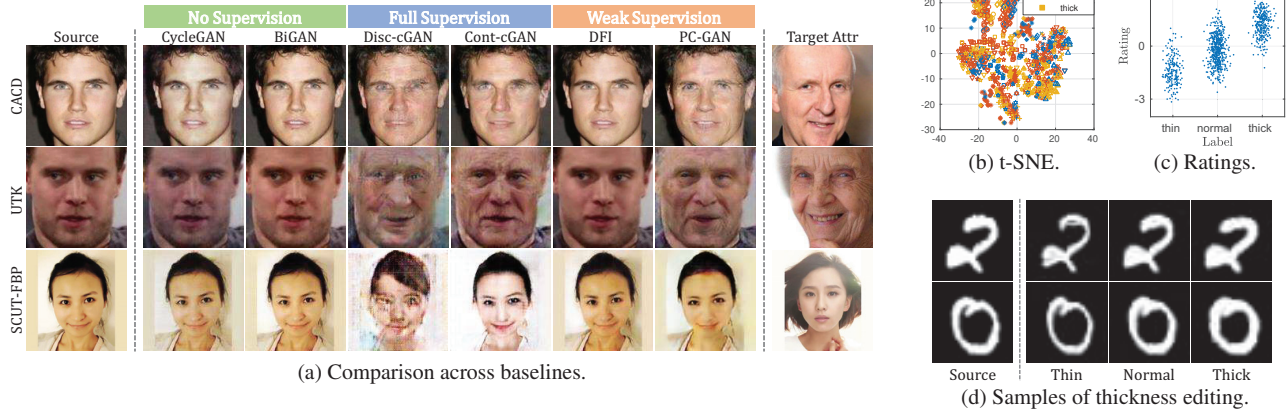
(a) Comparison across baselines.

(b) t-SNE.

(c) Ratings.

(d) Samples of thickness editing.

Figure 4: Results of facial attribute editing and Annotated MNIST: (**left a**) results of various baselines on different datasets, unsupervised baselines cannot effectively change the attribute intensity; (**right b-d**) results on Annotated MNIST. (b) t-SNE visualization of raw pixels, shapes correspond to numbers, and colors represent thickness levels. (c) Labels are jittered with random noise for better visualization.

| Dataset | Real | No Supervision | | Full Supervision | | Weak Supervision | |
		CycleGAN	BiGAN	Disc-CGAN	Cont-CGAN	DFI	PC-GAN
CACD	94.37(train) 49.00(val)	20.52	19.66	46.02	41.62	20.92	48.44
UTK	98.19(train) 76.80(val)	19.46	20.50	71.44	59.16	22.90	63.88
SCUT-FBP	100.00(train) 58.00(val)	19.75	20.38	29.63	46.25	22.69	40.00
Average Rank	–	5.67	5.33	2.00	2.33	4.00	**1.67**

Table 1: Evaluation of classification accuracies on synthesized images, higher is better.
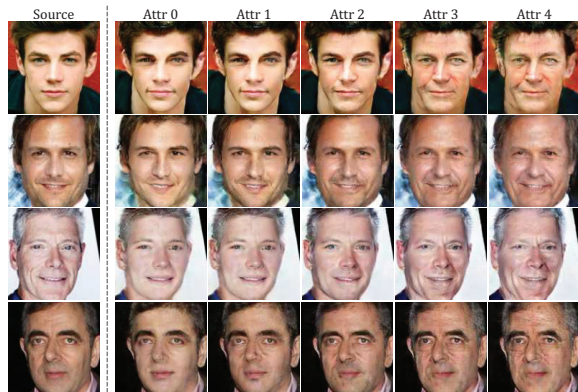

Figure 5: Results on CACD. The target attribute is age. Values from `Attr0` to `Attr4` correspond to age of 15, 25, 35, 45 and 55, respectively.
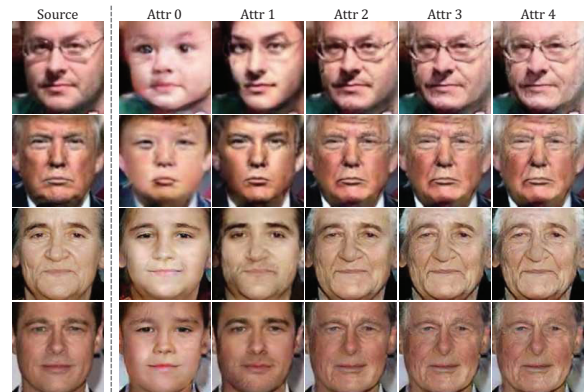

Figure 6: Results on UTKFace. The target attribute is age. Values from `Attr0` to `Attr4` correspond to age of 10, 30, 50, 70 and 90, respectively.

For the MNIST experiment, stroke thickness is the desired attribute. As illustrated in Figure 4-b, the thickness information is still entangled. But in Figure 4-c, the thickness is correctly disentangled from the rest attributes.

We use CACD and UTK for age progression, SCUT-FBP and CelebA for attractiveness experiment. Since no true relatively labeled dataset is publically available, pairs are *simulated* from "ground-truth" attribute intensity given in the dataset. The tie margins within which two candidates are considered equal are 10, 10, and 0.4 for CACD, UTK, and

SCUT-FBP, respectively. This also simplifies the quantitative evaluation process since one can directly measure the prediction error for absolute attribute intensities. Notice that CelebA only provides binary annotations, from which pairwise comparisons are simulated. Interestingly, the Elo rating network is still able to recover approximate ratings from those binary labels.

Furthermore, since CACD, UTKFace, SCUT-FBP, and CelebA are all human face dataset, we add an identity preserving loss term (Wang et al. 2018) to enforce identity
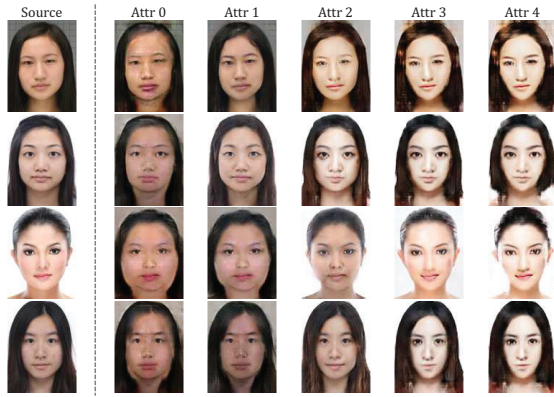
Figure 7: Results on SCUT-FBP. The target attribute is attractiveness score (1 to 5). Values from `Attr0` to `Attr4` correspond to score of 1.375, 2.125, 2.875, 3.625 and 4.5, respectively.
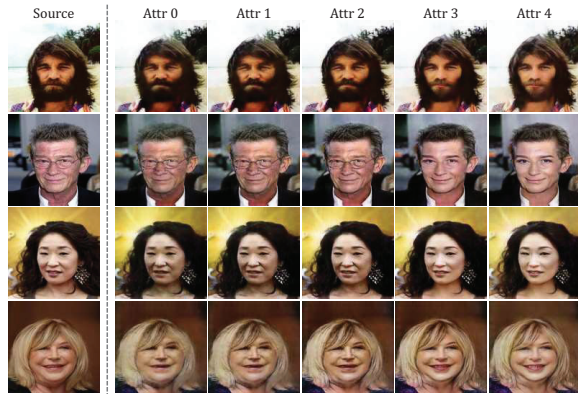

Figure 8: Results on CelebA. The target attribute is attractiveness. We take the cluster mean of ratings for "attractive" being -1 and 1 as `Attr0` and `Attr4` respectively. `Attr1` to `Attr3` are then linearly sampled. Results show a smooth transition of visual features, for example, facial hair, aging related features, smile lines, and shape of eyes.

preservation: $\mathcal{L}_{idt} = \mathbb{E}_{x \sim p(x), y \sim p(y)} \|h(\mathcal{G}(x,y)) - h(x)\|_2^2$. Here, $h(\cdot)$ denotes a pre-trained convnet.

Learning by Pairwise Comparison

Rating visualization. Figure 9 presents the predicted ratings learned from CACD, UTKFace, and SCUT-FBP from left to right. The ratings learned from pairwise comparisons highly correlate with the ground-truth labels, which indicates that the rating resembles the attribute intensity well. The uncertainties v.s. ground-truth labels is visualized in Figure 10. The plots show a general trend that the model is more certain about instances with extreme attribute values than those in the middle range, which matches our intuition. Additional attention-based visualizations are given in Supplementary.

Noise resistance. As mentioned previously, not only does pairwise comparison require less annotating effort, it tends to yield more accurate annotations. Consider a simple set-

	CACD			**UTKFace**		
Model	Acc (%)	IS	FID	Acc (%)	IS	FID
CNN-CGAN	35.04	2.14±0.02	31.08±0.54	40.12	2.69±0.03	26.58±0.51
BNN-CGAN	37.64	2.38±0.04	27.36±0.36	38.54	2.72±0.03	26.56±0.40
BNN-RCGAN	41.02	2.45±0.03	30.22±0.51	43.64	2.84±0.04	25.25±0.39

Table 2: Ablation study of Bayesian uncertainty estimation. `CNN-CGAN` is the normal non-Bayesian Elo rating network without uncertainty estimations; `BNN-CGAN` uses the average ratings for a single image; `BNN-RCGAN` is the full Bayesian model with a noise-robust CGAN.

Loss				**CACD**			**UTKFace**		
CGAN	rec	cyc	idt	Acc (%)	IS	FID	Acc (%)	IS	FID
✓	✓	✓	✓	48.08	2.87±0.04	27.90±0.44	62.74	3.50±0.04	21.63±0.52
✓	✗	✓	✓	39.50	2.93±0.04	25.68±0.46	56.90	3.38±0.05	24.98±0.88
✓	✓	✗	✓	50.86	3.10±0.04	25.93±0.55	60.56	3.39±0.05	23.70±0.65
✓	✓	✓	✗	48.60	3.05±0.03	26.81±0.59	63.92	3.60±0.05	27.65±0.75
✓	✗	✗	✓	48.98	3.01±0.03	26.90±0.67	66.34	3.65±0.04	25.39±0.86
✓	✗	✓	✗	24.28	3.06±0.04	24.01±0.66	50.42	3.02±0.04	48.80±1.70
✓	✓	✗	✗	43.86	2.94±0.05	24.27±0.58	62.42	3.54±0.04	32.87±1.47
✓	✗	✗	✗	20.08	1.59±0.02	293.03±1.40	34.88	2.16±0.04	187.98±2.17

Table 3: Ablation studies of different loss terms in CGAN training. CGAN represents $\mathcal{L}_{CGAN}$, rec represents $\mathcal{L}_{rec}$ and so on.

ting: if all annotators (annotating the absolute attribute value) exhibit the same random noise with a tie margin M, then the corresponding pairwise annotation with the same tie margin would absorb the noise. We provide an empirical study in the Supplementary.

Conditional Image Synthesis

Baselines. We consider two unsupervised baselines CycleGAN and BiGAN, two fully-supervised baselines Disc-CGAN and Cont-CGAN, and DFI in a similar weakly-supervised setting.

- **CycleGAN** (Zhu et al. 2017) learns an encoder (or a "generator" from images to attributes) and a generator between images and attributes simultaneously.
- **ALI/BiGAN** (Donahue, Krähenbühl, and Darrell 2016; Dumoulin et al. 2016) learns the encoder (an inverse mapping) with a single discriminator.
- **Disc-CGAN/IPCGAN** (Wang et al. 2018) takes discretized attribute intensities (one-hot embedding) as supervision.
- **Cont-CGAN** uses the same CGAN framework as PC-GAN but ratings are replaced by true labels. It is an upper bound of PC-GAN.
- **DFI** (Upchurch et al. 2017) can control the intensity of attribute intensity continuously, however, cannot change the intensity quantitatively. To transform x into $\tilde{x}'$, we assume $\phi(\tilde{x}') = \phi(x) + \alpha w$ and compute $y' = w \cdot \phi(x')$ (w is the attribute vector), then α is given by $\alpha = (y' - w \cdot \phi(x))/\|w\|_2^2$.

Qualitative results. In Figure 4, we compare our results with all baselines. For each row, we take a source and a target image as inputs and our goal is to edit the attribute value of the source image to be equal to that of the target image. PC-GAN is competitive with fully-supervised baselines while all unsupervised methods fail to change attribute intensities.
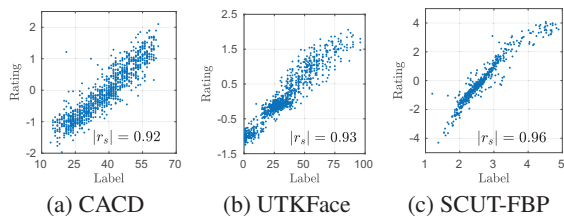
(a) CACD   (b) UTKFace   (c) SCUT-FBP

Figure 9: Visualization of learned ratings for different datasets. r_s denotes the Spearman's rank correlation coefficient.
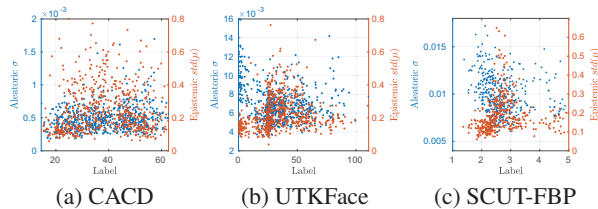

(a) CACD   (b) UTKFace   (c) SCUT-FBP

Figure 10: Visualization of the predictive uncertainty of learned ratings for different datasets (best viewed in color). Aleatoric (data-dependent) and epistemic (model-dependent) uncertainties are plotted separately.

More results are shown in Figure 5, 7, 6, where the target rating value is the average of (cluster mean) a batch of (10 to 50) labeled images. From Figure 5, we see aging characteristics like receding hairlines and wrinkles are well learned. Figure 6 shows convincing indications of rejuvenation and age progression. Figure 7 shows results for SCUT-FBP, which is inherently challenging because of the size of the dataset. Compared to datasets such as CACD, SCUT-FBP is significantly smaller, with only 500 images in total (from which we take 400 for training). Training on large datasets, as the CelebA experiment in Figure 8 shows, our model produces convincing results. We also find that the model is capable of learning important patterns that correspond to attractiveness, such as in the hairstyle and the shape of the cheek shown in Figure 7. (The result does not represent the authors' opinion of attractiveness, but only reflects the statistics of the annotations.)

Method	CACD		UTKFace	
	Quality (%)	Acc (%)	Quality (%)	Acc (%)
Real	97	36	88	52
PC-GAN	57	33	56	50
Cont-CGAN	60	31	55	37
Disc-CGAN	64	30	54	45

Table 4: AMT user studies. 100 images are sampled uniformly for each method with 20 images in each group.

Quantitative results. For quantitative evaluations, we report in Table 1 classification accuracy (Acc) evaluated on synthesized images. In our experiments, we train classifiers to predict attribute intensities of images into discrete groups (CACD 11-20, 21-30, up to > 50; UTK 1-20, 21-40, up to > 80, SCUT-FBP 1-1.75, 1.75-2.5, up to > 4).

PC-GAN demonstrates comparable performance with fully-supervised baselines and are significantly better than unsupervised methods. Additional metrics are reported in the Supplementary.

AMT user studies. We also conduct user study experiments. Workers from Amazon Mechanical Turk (AMT) are asked to rate the quality of each face (good or bad) and vote to which age group a given image belongs. Then we calculate the percentage of images rated as good and the classification accuracy. Table 4 shows that PC-GAN is on a par with the fully-supervised counterparts.

Ablation Studies

Supervision. The comparisons in Table 1 serve as an ablation study of full, no, and weak supervision, where PC-GAN is on a par with fully-supervised and significantly better than unsupervised baselines.

Uncertainty. The ablation study of the effectiveness of adding Bayesian uncertainties to achieve robust conditional adversarial training is given in Table 2. The three variants considered in the table differ in how much the Bayesian neural net is involved in the whole training pipeline: `CNN-CGAN` is a non-Bayesian Elo rating network plus a normal CGAN, `BNN-CGAN` learns a Bayesian encoder and yields the average ratings for a given image, and `BNN-RCGAN` trains a full Bayesian encoder with a noise-robust CGAN. Results confirm that the performance can be boosted by integrating an uncertainty-aware Elo rating network and an extended robust conditional GAN.

GAN loss terms. An ablation study of CGAN loss terms is provided in Table 3. Notice that setting some losses to zero is a special case of our full objective under different λs. Although we did not extensively tune λ's values since it is not the main focus of this paper, we conclude that $\mathcal{L}_{rec}$ is the most important term in terms of image qualities.

Conclusion

In this paper, we propose a noise-robust conditional GAN framework under weak supervision for image attribute editing. Our method can learn an attribute rating function and estimate the predictive uncertainties from pairwise comparisons, which requires less annotation effort. We show in extensive experiments that the proposed PC-GAN performs competitively with the supervised baselines and significantly outperforms the unsupervised baselines.

Acknowledgments

References

Bora, A.; Price, E.; and Dimakis, A. G. 2018. Ambientgan: Generative models from lossy measurements. *ICLR* 2:5.

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, 89–96. ACM.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, 2172–2180.

Chen, B.-C.; Chen, C.-S.; and Hsu, W. H. 2014. Cross-age reference coding for age-invariant face recognition and retrieval. In *European conference on computer vision*, 768–783. Springer.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

Dumoulin, V.; Belghazi, I.; Poole, B.; Mastropietro, O.; Lamb, A.; Arjovsky, M.; and Courville, A. 2016. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*.

Elo, A. E. 1978. *The rating of chessplayers, past and present*. Arco Pub.

Fürnkranz, J., and Hüllermeier, E. 2010. Preference learning and ranking by pairwise comparison. In *Preference learning*. Springer. 65–82.

Gal, Y., and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Han, L.; Murphy, R. F.; and Ramanan, D. 2018. Learning generative models of tissue organization with supervised gans. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 682–690. IEEE.

Herbrich, R.; Minka, T.; and Graepel, T. 2007. Trueskill™: a bayesian skill rating system. In *Advances in neural information processing systems*, 569–576.

Kendall, A., and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, 5574–5584.

Kim, B. 2017. Annotated mnist: Thickness and skew labeler for mnist handwritten digit dataset. https://github.com/1202kbs/Annotated_MNIST.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.; et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lucic, M.; Tschannen, M.; Ritter, M.; Zhai, X.; Bachem, O.; and Gelly, S. 2019. High-fidelity image generation with fewer labels. *arXiv preprint arXiv:1903.02271*.

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

Negahban, S.; Oh, S.; and Shah, D. 2016. Rank centrality: Ranking from pairwise comparisons. *Operations Research* 65(1):266–287.

Odena, A.; Olah, C.; and Shlens, J. 2016. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*.

Radinsky, K., and Ailon, N. 2011. Ranking from pairs and triplets: information quality, evaluation methods and query complexity. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 105–114. ACM.

Thekumparampil, K. K.; Khetan, A.; Lin, Z.; and Oh, S. 2018. Robustness of conditional gans to noisy labels. In *Advances in Neural Information Processing Systems*, 10271–10282.

Upchurch, P.; Gardner, J. R.; Pleiss, G.; Pless, R.; Snavely, N.; Bala, K.; and Weinberger, K. Q. 2017. Deep feature interpolation for image content changes. In *CVPR*, 6090–6099.

Wang, Z.; Tang, X.; Luo, W.; and Gao, S. 2018. Face aging with identity-preserved conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7939–7947.

Wauthier, F.; Jordan, M.; and Jojic, N. 2013. Efficient ranking from pairwise comparisons. In *International Conference on Machine Learning*, 109–117.

Xiao, F., and Jae Lee, Y. 2015. Discovering the spatial extent of relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, 1458–1466.

Xie, D.; Liang, L.; Jin, L.; Xu, J.; and Li, M. 2015. Scutfbp: A benchmark dataset for facial beauty perception. *arXiv preprint arXiv:1511.02459*.

Xu, Q.; Yang, Z.; Jiang, Y.; Cao, X.; Huang, Q.; and Yao, Y. 2019. Deep robust subjective visual property prediction in crowdsourcing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8993–9001.

Yan, S. 2016. Passive and active ranking from pairwise comparisons. Technical report, University of California, San Diego.

Zhang, Zhifei, S. Y., and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Zhou, Z.-H. 2017. A brief introduction to weakly supervised learning. *National Science Review* 5(1):44–53.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.