Supplementary Materials for Robust Conditional GAN from Uncertainty-Aware Pairwise Comparisons

Ligong Han[1], **Ruijiang Gao**[2], **Mun Kim**[1], **Xin Tao**[4], **Bo Liu**[3], **Dimitris Metaxas**[1]
[1]Department of Computer Science, Rutgers University
[2]McCombs School of Business, The University of Texas at Austin
[3]JD Finance America Corporation [4]Tencent YouTu Lab
l.han@rutgers.edu ruijiang@utexas.edu mun.kim@rutgers.edu
jiangsutx@gmail.com kfliubo@gmail.com dnm@cs.rutgers.edu

In Supplementary, we first show the analysis of CGAN loss terms and give a proof of Proposition 0.1. Then we provide an empirical study of how the number of pairs varies with the size of the dataset. The preliminary results on noise resistance is also presented. Next, we show qualitative attention visualization of the Elo rating network and report additional quantitative IS and FID scores for baselines and list details of network architectures. Finally, we show additional results on conditional image synthesis.

Analysis of Loss Terms

As a standard recall in (Goodfellow et al. 2014), the adversarial training results in minimizing the Jensen-Shannon divergence between the true conditional and the generated conditional. We show that the following proposition holds:

Proposition .1. *The global minimum of $\mathcal{L}(\mathcal{G},\mathcal{D})$ is achieved if and only if $q_{\mathcal{G}}(\tilde{x}'|x,y') = p_{\mathcal{E}}(\tilde{x}'|x,y')$, where p is the true distribution and $q_{\mathcal{G}}$ is the distribution induced by $\mathcal{G}$.*

Proof. (x', y') is sampled from true distribution, x is independently sampled and $\tilde{x}'$ is sampled from Generator $G(x, y')$, rewrite Equation 5 in integral form,

$$\begin{aligned}\mathcal{L}_{CGAN} = &\int p_{\mathcal{E}}(x',y')\log(\mathcal{D}(x',y'))dx'dy' + \\ &\int p(x)p_{\mathcal{E}}(y')q_{\mathcal{G}}(\tilde{x}'|x,y')\log(1-\mathcal{D}(\tilde{x}',y'))dxdy'd\tilde{x}' \\ = &\int p_{\mathcal{E}}(x,\tilde{x}',y')\log(\mathcal{D}(\tilde{x}',y')) + \\ &p_{\mathcal{E}}(x,y')q_{\mathcal{G}}(\tilde{x}'|x,y')\log(1-\mathcal{D}(\tilde{x}',y'))dxdy'd\tilde{x}',\end{aligned} \tag{1}$$

where we assume x and y' are sampled independently. We get the optimal discriminator $\mathcal{D}^*$ by applying Euler-Lagrange equation,

$$\mathcal{D}^* = \frac{p_{\mathcal{E}}(\tilde{x}'|x,y')}{p_{\mathcal{E}}(\tilde{x}'|x,y') + q_{\mathcal{G}}(\tilde{x}'|x,y')}. \tag{2}$$

Finally plugging $\mathcal{D}^*$ in $\mathcal{L}_{CGAN}$ yields,

$$\begin{aligned}\mathcal{L}_{CGAN}(\mathcal{G},\mathcal{D}^*) = &-2\log 2 \quad + \\ &2\int p_{\mathcal{E}}(x,y')\text{JSD}(p_{\mathcal{E}}(\tilde{x}'|x,y')||q_{\mathcal{G}}(\tilde{x}'|x,y'))\,dxdy',\end{aligned} \tag{3}$$

where JSD is the Jensen-Shannon divergence. Since JSD is always non-negative and reaches its minimum if and only if $q_{\mathcal{G}}(\tilde{x}'|x,y') = p_{\mathcal{E}}(\tilde{x}'|x,y')$ for $(x,y') \in \{(x,y') : p_{\mathcal{E}}(x,y') > 0\}$, $\mathcal{G}$ recovers the true conditional distribution $p_{\mathcal{E}}(\tilde{x}'|x,y')$ when $\mathcal{D}$ and $\mathcal{G}$ are trained optimally.

In addition, the reconstruction loss $\mathcal{L}^y_{rec}$, cycle loss $\mathcal{L}_{cyc}$, and identity preserving loss $\mathcal{L}_{idt}$ are all non-negative. Minimizing these losses will keep the equilibrium of $\mathcal{L}_{CGAN}$. If the encoder $p_{\mathcal{E}}(y|x)$ and the feature extractor $h(\cdot)$ are trained properly, $\mathcal{L}(\mathcal{G},\mathcal{D}^*)$ achieves its minimum when $\mathcal{G}$ is optimally trained. □

Proof of Proposition 0.1

Proof. For $\forall u, v \in V$, we define $\pi(u,v) = 1$ if $u < v$ and 0 otherwise, $w(u,v)$ measures the extent to which u should be prefered over v,

For any pair u, v, let

$$L_{u,v} = \pi(u,v)w(u,v) + \pi(v,u)w(v,u) \tag{4}$$

where $\pi(u,v)$ is the ground-truth and $w(v,u)$ is prediction from Elo ranking network.

Define

$$L = \Sigma_{u<v,u,v\in V} L_{u,v} \tag{5}$$

as our loss function and from results in (Radinsky and Ailon 2011), we have the lemma:

Lemma .1. *For $\delta > 0$, any $0 < \lambda < 1$, if we sample dn/λ^2 pairs uniformly with repetition from $\binom{V}{2}$, with probability $1-\delta$,*

$$L(V,w,\hat{\pi}) \le \lambda\left[\frac{c}{\sqrt{d}} + \sqrt{\frac{\log\frac{1}{\delta}}{dn}}\right]\binom{n}{2}. \tag{6}$$

Define

$$t = \lambda\left[\frac{c}{\sqrt{d}} + \sqrt{\frac{\log\frac{1}{\delta}}{dn}}\right]\binom{n}{2}, \tag{7}$$

and let $\delta = 1$, we get t_1 and $\mathbb{P}(L(\hat{\pi}) > t_1) \le \delta = 1$

$$t_1 = \lambda\left[\frac{c}{\sqrt{d}} + \sqrt{\frac{\log 1}{dn}}\right]\binom{n}{2}. \tag{8}$$

$$\mathbb{E}(L(\hat{\pi})) = \int_0^\infty \mathbb{P}(L(\hat{\pi}) > t)dt \leq t_1 + \int_{t_1}^\infty \mathbb{P}(L(\hat{\pi}) > t)dt \tag{9}$$

From Equation 7,

$$\delta = exp(-\frac{1}{2}\sigma_n^2(t-\mu_n)^2) \tag{10}$$

where $\sigma_n^2 = \frac{\lambda^2(n(n-1))^2}{4dn}$, $\mu_n = \frac{\lambda n(n-1)c}{2\sqrt{d}}$.

Plugging back in Equation 9,

$$\mathbb{E}(L(\hat{\pi}) \leq t_1 + \sqrt{2\pi\sigma_n^2}$$

$$= \lambda\left[\frac{c}{\sqrt{d}} + \sqrt{\frac{\log 1}{dn}}\right]\binom{n}{2} + 2\lambda\sqrt{2\pi}\frac{n(n-1)}{s\sqrt{dn}}$$

$$= \lambda\left[\frac{c}{\sqrt{d}} + \frac{\sqrt{\log 1} + \sqrt{8\pi}}{\sqrt{dn}}\right]\binom{n}{2}. \tag{11}$$

Set $d = 16c^2$, for $\lambda/4 > \epsilon_0 > 0$, there is n_0 so that if $n > n_0$,

$$\mathbb{E}(L(\hat{\pi})) \leq (\lambda/4 + \epsilon_0)\binom{n}{2} \leq \lambda/2\binom{n}{2}. \tag{12}$$

□

Number of Pairs

To experimentally verify the number of pairs needed to learn a rating, we sampled from UTKFace (Zhang and Qi 2017) subsets of sizes 100, 500, 1000, 2000, 5000 and 10000, and train Elo rating networks with different number of pairs for each subset. As illustrated in Figure 1, to achieve a Spearman correlation above 0.9, approximately $2n$ pairs are needed, where n is the size of the subset. $n \log n$ comparisons are needed for exact recovery of ranking between n objects. Through our ranking network, we need $\mathcal{O}(n)$ comparisons to learn rating that is close enough to the true attribute strength and also keeping the space between objects. Annotation of absolute attribute strength is very noisy and usually takes $\mathcal{O}(n)$ annotations because of majority voting (e.g. $3n$ if 3 workers per instance), our method doesn't require more effort in annotation and pairwise comparisons are easier to annotate comparing to absolute attribute strength, which will lead to a faster finishing time in crowd-sourcing phase.

Noise Resistance

Considering there is noise when annotating the absolute labels. Taking age annotation as an example, we assume annotators will give x an age $\Omega'(x)$ that deviates from the true age $\Omega(x)$ by a random noise: $\Omega'(x) = \Omega(x) + z, z \sim$ Unif $\left(-\frac{M}{2}, \frac{M}{2}\right)$, where M is the tie margin in Figure 2. As shown, the correlation curve of ratings drops slowly until the noise level is too high. Although only the curve on SCUT-FBP shows superior results over the ground-truth label, the general trend is that the rating curves decrease slower than the absolute label curves. This demonstrates the Elo rating network's potential of noise resistance.

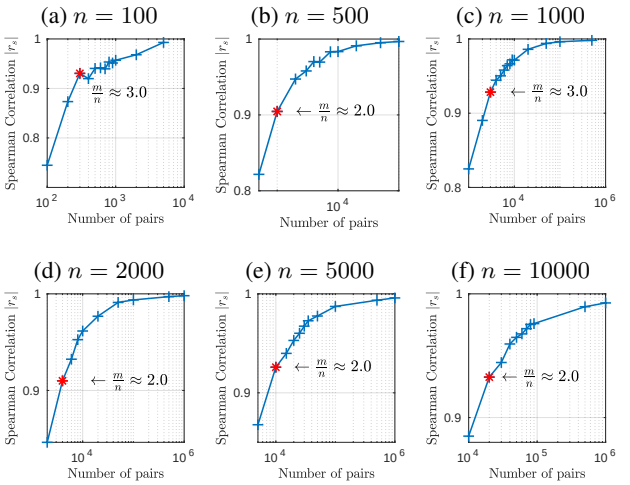

Figure 1: Number of pairs m v.s. Spearman correlation r_s. Different subsets of images (of number $n = 100, \ldots, 10000$) are randomly selected from the UTKFace dataset. For each subset, different number of pairs (denoted by m) are randomly sampled. The smallest number of pairs with a Spearman's rank correlation coefficient that exceeds 0.9 is marked by a red asterisk symbol $*$. To achieve high correlations between ratings and labels (in terms of $|r_s| \geq 0.9$), approximately $2n$ pairs are required.

We choose UTKFace dataset to investigate how conditional synthesis results might be affected by margins. In Table 1, Spearman correlations and Inception Scores evaluated on UTKFace under different margin values are reported.

Margin	**Corr**	**Acc (%)**	**IS**
5	0.93	73.26	3.70±0.07
15	0.91	64.18	3.56±0.06
25	0.88	73.26	3.78±0.04
35	0.85	60.74	3.50±0.06

Table 1: Spearman correlations (**Corr**), Inception Scores (**IS**) evaluated on UTKFace under different margin values. Pairs are randomly sampled and CGANs are trained using different pairs.

Attention Visualization

The proposed Elo rating network is visualized using Grad-CAM (Selvaraju et al. 2017). In Figure 3-a, local regions that are critical for decision making are highlighted: for CACD and UTKFace, aging indicators such as forehead wrinkles, crow's feet eyes (babies usually have big eyes) are highlighted; for SCUT-FBP, the gradient map highlights facial regions like eyes, nose, pimples etc. Similar to DFI, if viewing the rating as deep features, one can optimize over the input image to obtain a new image with desired attribute intensity. We thus invert the encoders to see what a "typical" image with extreme attribute intensity would look like by optimizing the average face as shown in Figure 3-b.

(a) **Inception Score** (higher is better)

		weak supervision		full supervision		no supervision	
Dataset	Real	PC-GAN	DFI	Cont-CGAN	Disc-CGAN	CycleGAN	BiGAN
CACD	3.89 ± 0.05	2.89 ± 0.06	3.35 ± 0.06	2.85 ± 0.03	2.95 ± 0.04	2.96 ± 0.03	3.27 ± 0.04
UTK	4.29 ± 0.05	3.55 ± 0.06	3.26 ± 0.06	3.52 ± 0.04	3.66 ± 0.04	3.09 ± 0.06	3.20 ± 0.06
SCUT-FBP	4.20 ± 0.05	2.88 ± 0.11	2.93 ± 0.07	2.39 ± 0.14	1.37 ± 0.02	2.85 ± 0.15	3.05 ± 0.15

(b) **Fréchet Inception Distance** (lower is better)

	weak supervision		full supervision		no supervision	
Dataset	PC-GAN	DFI	Cont-CGAN	Disc-CGAN	CycleGAN	BiGAN
CACD	28.20 ± 0.65	25.18 ± 0.73	28.53 ± 0.72	28.13 ± 0.71	26.76 ± 0.64	24.69 ± 0.62
UTK	24.86 ± 0.84	28.32 ± 0.75	28.42 ± 0.98	33.26 ± 1.49	23.16 ± 0.75	19.72 ± 0.79
SCUT-FBP	97.21 ± 2.81	48.67 ± 1.42	114.89 ± 3.08	188.09 ± 3.91	87.07 ± 3.21	81.16 ± 2.93

Table 2: Inception Scores (**IS**) and Fréchet Inception Distances (**FID**). IS and FID are computed from 20 splits with 1000 images in each split. Unsupervised baselines fail to edit source images to a desired attribute strength and show classification accuracies close to a random guess (around 20%), however, they have misleadingly high IS and low FID scores (because changes are subtle compared to the source images).

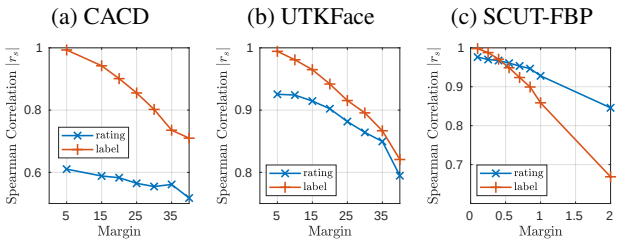

Figure 2: Noise resistance. Spearman correlations between ground-truth labels and ratings or noisy labels under different tie margins (a tie margin is the range within which an agent is indifferent between two alternatives).

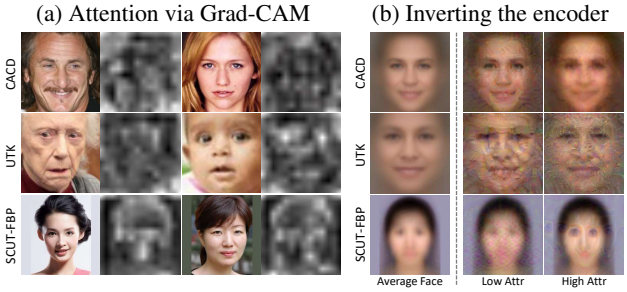

Figure 3: (a) Attention visualization for Elo rating network via Grad-CAM. (b) Inverting the Elo rating network by optimization over the input image (average faces) to match low/high attribute intensity.

IS and FID Scores

Additional Inception Scores (IS) (Salimans et al. 2016), Fréchet Inception Distances (FID) (Heusel et al. 2017) are reported in Table 2. Classifiers for evaluating classification accuracies are also used to compute Inception Scores and as auxiliary classifiers in training Disc-CGAN/IPCGAN. The unsupervised baselines have high Inception Scores and low Fréchet Inception Distances but very low classification accuracies since their outputs are almost identical to source images. Collectively, PC-GAN demonstrates comparable performance with fully-supervised baselines and are significantly better than unsupervised methods.

Network Architectures

We show the architectures of our Elo ranking network as well as the spatial transformer network in Table 3. Facial attribute classifiers are finetuned ResNet-18 (He et al. 2016).

Layers	Weights	Activations
Input image		$224 \times 224 \times 3$
`ResNet-18 features`		$7 \times 7 \times 512$
`conv, pad 1, stride 1`	$3 \times 3 \times 64$	$7 \times 7 \times 64$
`BatchNorm, LeakyReLU`		
`conv, pad 1, stride 1`	$3 \times 3 \times 1$	$7 \times 7 \times 1$
`Global AvgPool`		$1 \times 1 \times 1$

Table 3: Architecture of Elo ranking network. `ResNet-18 features` are the CNN layers before its classifier.

Additional Results

Additional results of our PC-GAN and two fully-supervised baselines Cont-CGAN and Disc-CGAN/IPCGAN (Wang et al. 2018) on CACD, UTKFace, and SCUT-FBP datasets are given in Figure 4, 5, and 6 respectively. Results for unsupervised baselines are not shown since the changes in outputs are subtle. For CACD, attribute values (from `Attr0` to `Attr4`) correspond to ages of 15, 25, 35, 45 and 55; for UTK, attribute values correspond to ages of 10, 30, 50, 70 and 90; for SCUT-FBP, attribute values correspond to scores of 1.375, 2.125, 2.875, 3.625 and 4.5, respectively.

PC-GAN, Cont-CGAN and Disc-CGAN perform similarly on CACD. Disc-GAN performs much worse on UTKFace and SCUT-FBP, presumably due to the discretization of attribute strength. For example, in SCUT-FBP, the number of images are unevenly distributed across discretized attribute groups, that is, groups with least and largest attribute strength (attractiveness) have only limited images. In this case, we are more likely to see mode collapse in Disc-CGAN. As a result, Disc-CGAN is outputting same images for `Attr0` and `Attr4` in Figure 8. PC-GAN and Cont-CGAN have a similar quality in synthesized images in all three datasets, which shows PC-GAN can synthesize images of same qualities using pairwise comparisons.

References

[Goodfellow et al. 2014] 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.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[Heusel et al. 2017] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 6626–6637.

[Radinsky and Ailon 2011] 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.

[Salimans et al. 2016] Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2234–2242.

[Selvaraju et al. 2017] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

[Wang et al. 2018] 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.

[Zhang and Qi 2017] 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.

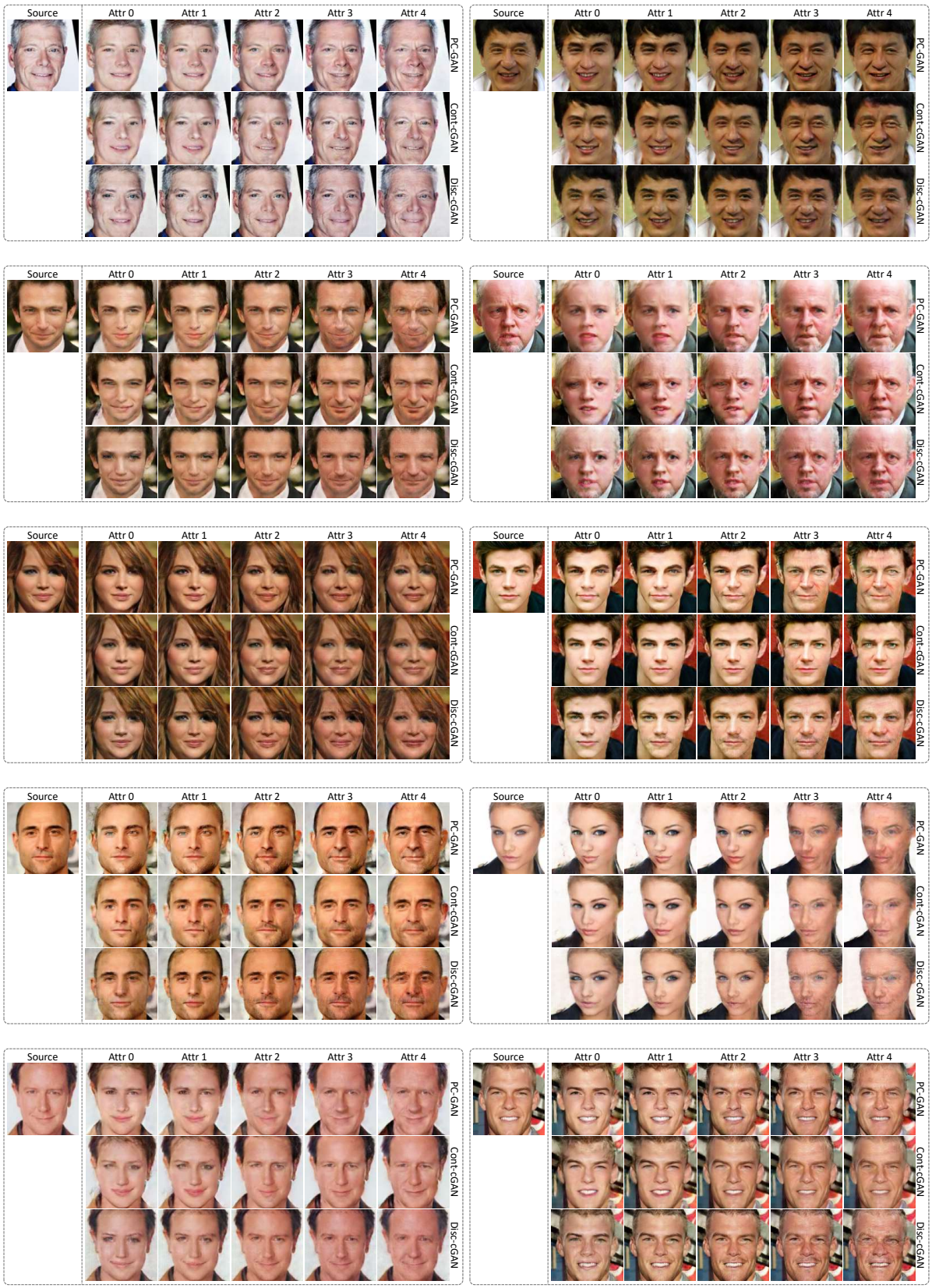

Figure 4: Comparison of PC-GAN with Cont-CGAN and Disc-CGAN on the CACD dataset. Attribute values from `Attr0` to `Attr4` correspond to age of 15, 25, 35, 45 and 55, respectively.

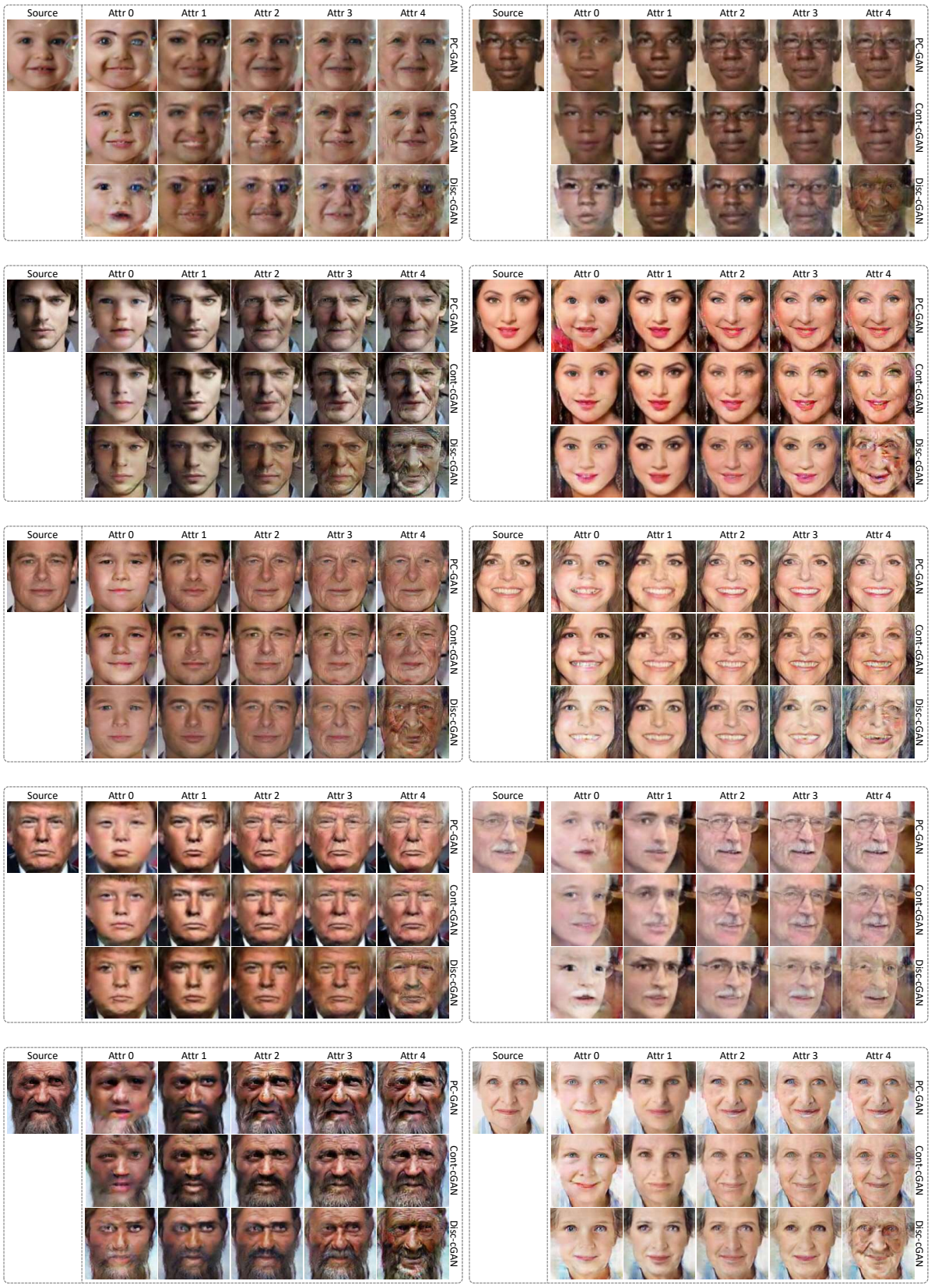

Figure 5: Comparison of PC-GAN with Cont-CGAN and Disc-CGAN on the UTKFace dataset. Attribute values from `Attr0` to `Attr4` correspond to age of 10, 30, 50, 70 and 90, respectively.

Figure 6: Comparison of PC-GAN with Cont-CGAN and Disc-CGAN on the SCUT-FBP dataset. Attribute values from `Attr0` to `Attr4` correspond to score of 1.375, 2.125, 2.875, 3.625 and 4.5, respectively.