Vol.2018-CG-172 No.11
Vol.2018-DCC-20 No.11
Vol.2018-CVIM-214 No.11
2018/11/8

標準的な顔画像データセットを用いた顔方位変換のための深層生成モデル

森川 将平[1,a)]　齋藤 豪[1,b)]

概要：人物の顔画像に対する顔方位変換画像の生成は顔認識の分野において非常に重要な課題の一つである．近年提案されている深層学習を用いた顔方位変換画像の生成モデルでは，一人につき複数枚用意された多様な人物の画像と人物ラベルが付与されたデータセットを学習に必要とし，ある環境下で様々な方位から撮影された人物の顔画像データセットを用いている．本研究ではそのようなデータセットを用いずに顔方位変換を学習する生成モデルを提案する．

Deep Face Rotation with Ordinary Dataset

Shohei Morikawa[1,a)]　Suguru Saito[1,b)]

1. はじめに

人物の顔画像に対して顔方位変換を施した画像の生成は顔認識の分野において非常に重要な課題の一つである．近年，深層学習の一つである畳み込みニューラルネットワークによって画像生成の分野には著しい発展があり，この問題も例外ではない．先行研究では顔方位変換を行うネットワークを学習するためのデータセットとして Multi-PIE [5] を用いている．このデータセットは多視点から人物を撮影した顔画像によって作成され，人物に対してそれぞれ固有のラベルが割り当てられているため顔認識タスクにおいてしばしば用いられる．また顔方位変換画像の生成においても，顔画像に対する方位変換後の目的画像が存在するという点で非常に優秀である．しかしこのようなデータセットはデータの作成や追加拡張が非常に困難であり，一般的に利用可能なデータセットであるとはいえない．そこで我々は先行研究には提案されない標準的な顔画像データセットのみを用いて学習された顔方位変換画像の生成モデルを提案し，方位変換後の目的画像が存在しない学習においても人物の個性が保存された顔方位変換画像の生成を目指す．ここで標準的な顔画像データセットとは，CelebA [17] や LFW [9] のように制約のない撮影条件の下で撮影された顔写真によって構成されたデータセットを指す．

[1] 東京工業大学　情報理工学院
School of Computing, Tokyo Institute of Technology
[a)] shohei@img.cs.titech.ac.jp
[b)] suguru@c.titech.ac.jp

2. 関連研究

2.1 深層生成モデル

現在，画像処理の分野において深層学習を用いた手法は数多く提案され，その分野の発展に多大な貢献がある．画像生成においては Variational Auto-Encoder (VAE) [14] や Generative Adversarial Networks (GAN) [4], [22] のような生成モデルと呼ばれる手法が一般的に用いられる．

VAE は従来の自己符号化器によって得られる特徴空間に対して確率分布を仮定したものであり，学習によってデータの分布をモデル化することで仮定した分布のサンプリングからデータセットに存在しない新しいデータサンプルを獲得することが可能である．GAN は生成ネットワークと識別ネットワークの 2 つのネットワークから構成され，生成ネットワークがある分布からのサンプリングされたノイズからデータセットに存在する様なデータを生成し，識別ネットワークはデータが訓練データであるか生成ネットワークによって生成されたデータであるかを正しく識別するように学習する．このとき生成ネットワークが識別ネットワークの分別が困難になるようなデータを生成するよう

Vol.2018-CG-172 No.11
Vol.2018-DCC-20 No.11
Vol.2018-CVIM-214 No.11
2018/11/8

に二者間のミニマックスゲームによって学習することで，最終的に本物と区別のつかないデータを生成するネットワークの獲得が期待できる．

これらの生成モデルを基礎として条件を設けたデータ生成を試みる条件付き生成モデルも提案され [2], [7], [13], [20]，顔方位変換画像の生成のために顔方位を条件とした条件付き生成モデルを提案モデルに導入する．また生成モデル VAE と GAN を組み合わせたモデルも数多く提案されており [1], [15], [18], [19]，我々の手法も 2 つの生成モデルを組み合わせた構造を持つ．

2.2 顔方位変換画像の生成

深層ニューラルネットワークを用いて人物の顔画像からその顔の向きが変換された画像を生成する手法は数多く提案されている．

Yim ら [29] は入力画像の顔方位変換の過程を複数のタスクに分割し，それぞれの処理について畳み込みニューラルネットを用いることで画像を合成する．対照的に Zhu ら [32] は顔画像の個性表現と方位表現を異なるニューロンで処理することで 2 つの表現を分離したシングルタスクによる顔方位変換手法を提案している．その他にも自己符号化器を用いて段階的に方位変換を学習することで最終的な方位変換を目指す手法では，複数の符号化器を用いる Kan らの手法 [12] や，再帰型ニューラルネットワークへ拡張して反復的な方位の修正を行う Yang らの手法 [28] がある．

また深層生成モデルを用いた手法もいくつか提案されている [10], [26], [27], [30], [31]．これらの生成モデルは全て GAN を利用した手法で，[26], [27], [30] では入力画像に対して任意の方位変換に対応した画像生成モデルが提案され，Zheng ら [31] はいくつかの表情についても制御可能な顔方位変換モデルを提案する．

ここに挙げた全ての手法は Multi-PIE [5] に代表されるように人物を多視点から撮影し，顔方位に関して制御された環境下で作成されたデータセットを用いている．このデータセットは各人物に対してそれぞれ固有のラベルが付与され複数視点における顔認識タスクにおいてしばしば利用される．また顔方位変換画像の生成においても，ある固有の人物に対して方位変換後の目的画像が存在するという点で学習用データとして非常に優秀である．しかし Multi-PIE データセットに含まれる被験者の数は高々 337 人であり，多様な顔表現の学習においてその数が十分であるかは不明である．多様な顔画像群から同一人物の顔画像を探索しそれぞれに固有ラベルを割り振ることや，顔方位に関して同様に制御された画像をこのデータセットの新しいデータとして追加拡張することは非常に困難であり，一般的に利用可能なデータセットであるとはいえない．

我々の貢献はこのような特殊な環境条件において作成されたデータセットを用いずに，方位変換後の目的画像が存在しない標準的な顔画像データセットを用いて人物の個性表現を学習し，顔方位変換を行う生成モデルの学習フレームワークを提案することである．

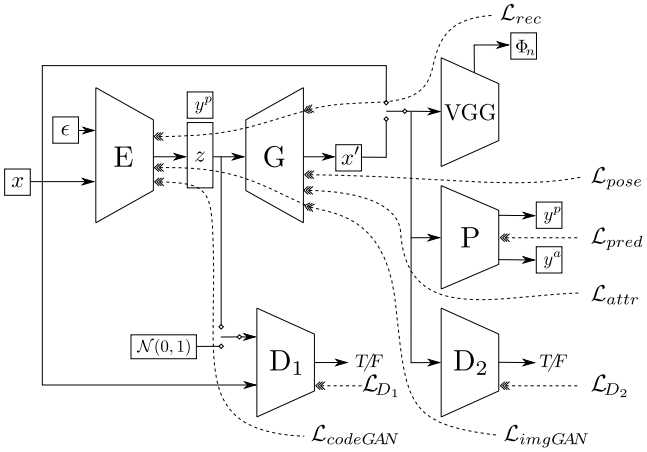

図 1 ネットワーク概要
Fig. 1 The architecture of our network

3. 提案手法

3.1 ネットワーク

我々の提案するネットワークは Conditional VAE [13] と GAN [4], [22] を基本構造に持つ．ネットワークの概要を図 1 に示す．ネットワークは符号化ネットワーク，合成ネットワーク，推論ネットワーク，2 つの識別ネットワーク，そして VGG16 ネットワーク [24] の計 6 つのサブネットワークによって構成され，これらのネットワークを E，G，P，D_1，D_2，VGG で表す．E，G，P，D_2 はストライド 2 の畳み込み層が 1 層とストライド 2 の残差ブロック [6] が 3 層，そして 2 層の全結合層から構成される．D_1 は Adversarial Variational Bayes (AVB) [19] によって提案される識別ネットワークを用い，前述の構成に加えて潜在変数のための 3 層の全結合層を追加する．サブネットワークと残差ブロックの構成は表 1 に示される．E，G，P の中間層に用いられる全ての畳み込みは Inception モジュール [25] に置き換えられ，中間層の出力には Batch Normalization [11] を適用する．また D_1，D_2 には全ての層に Spectral Normalization [21] を適用する．VGG は E，G を訓練するために ImageNet データセット [3] によって事前学習された VGG16 ネットワークが用いられ，学習によって重みは更新されない．P，D_1，D_2 は第一全結合層の代わりに Global Average Pooling (GAP) [16] を用いる．P 以外の全てのサブネットワークの活性化関数に Lealy ReLU，P の活性化関数に ReLU を用いる．

3.2 損失関数

VAE の学習に用いる再構成誤差 L_{rec} には Deep Feature Consistent VAE [8] で提案される VGG 損失を用いる。一

Vol.2018-CG-172 No.11
Vol.2018-DCC-20 No.11
Vol.2018-CVIM-214 No.11
2018/11/8

表 1　各サブネットワークと残差ブロックの構造
Table 1 The structure of sub-networks and the res block module

Encoder		
Layer	Filter/Stride	Output Size
Input	(Image+Noise)	$96 \times 96 \times (3+1)$
Conv1	$3 \times 3/1$	$96 \times 96 \times 32$
Res Block2	$3 \times 3/2$	$48 \times 48 \times 64$
Res Block3	$3 \times 3/2$	$24 \times 24 \times 128$
Res Block4	$3 \times 3/2$	$12 \times 12 \times 256$
FC1		256
FC2		128

Generator		
Layer	Filter/Stride	Output Size
Input	(Latent+Pose code)	$128+3$
FC1		256
FC2		$12 \times 12 \times 256$
Res Block1	$3 \times 3/\frac{1}{2}$	$24 \times 24 \times 128$
Res Block2	$3 \times 3/\frac{1}{2}$	$48 \times 48 \times 64$
Res Block3	$3 \times 3/\frac{1}{2}$	$96 \times 96 \times 32$
Conv4	$3 \times 3/1$	$96 \times 96 \times 3$

Predictor		
Layer	Filter/Stride	Output Size
Input	(Image)	$96 \times 96 \times 3$
Conv1	$3 \times 3/1$	$96 \times 96 \times 32$
Res Block2	$3 \times 3/2$	$48 \times 48 \times 64$
Res Block3	$3 \times 3/2$	$24 \times 24 \times 128$
Res Block4	$3 \times 3/2$	$12 \times 12 \times 256$
GAP		256
FC		$3+40$

Discriminator1 (for latents)		
Layer	Filter/Stride	Output Size
Input	(Image&Latent)	$96 \times 96 \times 3, 128$
Conv1,-	$3 \times 3/1$	$96 \times 96 \times 32, 128$
Res Block2,-	$3 \times 3/2$	$48 \times 48 \times 64, 128$
Res Block3, FC1	$3 \times 3/2$	$24 \times 24 \times 128, 256$
Res Block4, FC2	$3 \times 3/2$	$12 \times 12 \times 256, 256$
GAP, FC3		$256, 256$
Inner product		1

Discrimator2 (for images)		
Layer	Filter/Stride	Output Size
Input	(Image)	$96 \times 96 \times 3$
Conv1	$3 \times 3/1$	$96 \times 96 \times 32$
Res Block2	$3 \times 3/2$	$48 \times 48 \times 64$
Res Block3	$3 \times 3/2$	$24 \times 24 \times 128$
Res Block4	$3 \times 3/2$	$12 \times 12 \times 256$
GAP		256
FC		1

Residual Block		
Layer	Filter/Stride	Output Size
Input		$W \times H \times C^l$
Conv1A	$1 \times 1/n$	$\frac{W}{n} \times \frac{H}{n} \times C^{l+1}$
Conv1B	$3 \times 3/n$	$\frac{W}{n} \times \frac{H}{n} \times C^{l+1}$
Batch Norm		
Activation		
Conv2B	$1 \times 1/1$	$\frac{W}{n} \times \frac{H}{n} \times C^{l+1}$
add	(Conv1A+Conv2B)	$\frac{W}{n} \times \frac{H}{n} \times C^{l+1}$

般的に再構成誤差は入力画像と再構成された出力画像の $L2$ ノルムを用いるが，ここでは入出力画像を VGG へ入力することによって得られる中間特徴の $L2$ ノルムによって再構成誤差を計算する．VGG 損失は従来の VAE の生成結果に比べて G がより鮮明な画像を生成することが期待される．中間特徴には VGG16 ネットワークにおける Conv1_1，Conv2_1，Conv3_1 の 3 層の出力を用いている．

$$L_{rec} = \sum_n ||\Phi_n(x) - \Phi_n(G(z_x, y_x^p))||_2, \tag{1}$$

ここで $\Phi_n(x)$ は画像 x を VGG16 へ入力した際の中間畳み込み層 Conv n_1 の出力であり，$G(z_x, y_x^p)$ は画像 x の符号化によって得られる潜在変数 z_x と画像 x の持つ顔方位情報 y_x^p によって合成された画像である．

L_{D_1} と L_{D_2} は 2 つの識別ネットワーク D_1，D_2 を訓練するための損失である．D_1 は AVB [19] で提案される識別ネットワークである．従来の VAE では潜在空間の確率分布をモデル化するために事前分布と潜在空間の分布の 2 つの確率分布間の距離を Kullback-Leibler divergence によって E を訓練する．AVB では事前分布からのサンプリングであるか E による画像の符号化によって得られた潜在変数であるかを識別するように D_1 を訓練し，E が D_1 を騙すように敵対的損失 $L_{codeGAN}$ を用いて学習することで VAE における潜在空間の確率分布をモデル化する．D_2 は従来の GAN で提案される識別ネットワークである．D_2 は訓練データからサンプリングされた画像であるか G によって合成された画像であるかを識別し，G が D_2 を騙すように敵対的損失 L_{imgGAN} を用いて学習することで G がデータセットの画像と見分けが付かない画像を合成する．

$$\begin{aligned} L_{D_1} &= -\mathbb{E}_{z\sim p_z(z)}[logD_1(z)] \\ &\quad - \mathbb{E}_{x\sim p_{data}(x)}[log(1 - D_1(z_x))] \end{aligned} \tag{2}$$

$$\begin{aligned} L_{D_2} &= -\mathbb{E}_{x\sim p_{data}(x)}[logD_2(x)] \\ &\quad - \mathbb{E}_{z\sim p_z(z)}[log(1 - D_2(G(z, y^p)))] \end{aligned} \tag{3}$$

$$L_{codeGAN} = -\mathbb{E}_{x\sim p_{data}(x)}[logD_1(z_x)] \tag{4}$$

$$L_{imgGAN} = -\mathbb{E}_{z\sim p_z(z)}[logD_2(G(z, y^p))] \tag{5}$$

ここでは潜在空間の事前分布 p_z として多次元正規分布 $\mathcal{N}(0, I)$ を用いる．

推論ネットワーク P は入力された画像の持つ顔方位情報と属性情報を推定する．P は G の学習のために後述される条件損失と属性損失を算出するために L_{pred} によって訓練される．L_{pred} は顔方位の連続量推定と二項分類される属性のマルチラベル推定のために $L2$ ノルムと交差エントロピーを用いる．

$$\begin{aligned} L_{pred} &= ||y_x^p - P_{pose}(x)||_2 - y_x^a logP_{attr}(x) \\ &\quad - (1 - y_x^a)log(1 - P_{attr}(x)) \end{aligned} \tag{6}$$

ここで y_x^a は画像 x が持つ属性ラベルを表す．

G が入力された方位条件 y^p を考慮して画像生成を行うために条件損失 L_{pose} を用いる．L_{pose} は G への入力方位条件とその生成画像から P によって推定された方位情報の $L2$ ノルムによって定義される．

$$L_{pose} = ||y^p - P_{pose}(G(z, y^p))||_2. \tag{7}$$

ある潜在変数から G によって生成される人物画像が入力された方位条件 y^p の変化に対して不変であるように属性損失 L_{attr} を用いる．ある人物が顔方位を変化しても属性情報は変化しないと仮定し，L_{attr} は P の入力画像に対する属性情報の推定値と合成画像に対する属性情報の推定値が一致するように交差エントロピーによって定義される．

Vol.2018-CG-172 No.11
Vol.2018-DCC-20 No.11
Vol.2018-CVIM-214 No.11
2018/11/8

$$L_{attr} = -P_{attr}(x)logP_{attr}(G(z_x, y^p)) \\ -(1-P_{attr}(x)))log(1-P_{attr}(G(z_x, y^p))) \quad (8)$$

Algorithm 1 Training process

Require: m, the batch size. θ_X, initial X network parameters. ϵ is random noise. $\lambda_1 = 0.00003$. $\lambda_2 = 0.001$.

while θ_G has not converged **do**
  Sample$\{x, y^p_x, y^a_x\} \sim P_{data}$ a batch from the dataset;
  Get L_{pred} by Eq.6
  $z_x \leftarrow E(x, \epsilon)$
  Get L_{rec}, $L_{codeGAN}$, L_{attr} by Eq.1,4,8
  Sample$\{z\} \sim P_z$ a batch of random noise, and y^p same label as y^p_x;
  Get L_{D_1}, L_{pose} by Eq.2,7
  if Pre-training **then**
    L_{D_2}, $L_{imgGAN} \leftarrow$ Const.
  else
    Get L_{D_2}, L_{imgGAN} by Eq.3,5
  end if
  $\theta_P \overset{+}{\leftarrow} -\nabla_{\theta_P}(L_{pred})$
  $\theta_{D_1} \overset{+}{\leftarrow} -\nabla_{\theta_{D_1}}(L_{D_1})$
  $\theta_{D_2} \overset{+}{\leftarrow} -\nabla_{\theta_{D_2}}(L_{D_2})$
  $\theta_{E,G} \overset{+}{\leftarrow} -\nabla_{\theta_{E,G}}(\lambda_1 L_{rec} + L_{codeGAN} + \lambda_2 L_{imgGAN} + L_{pose} + L_{attr})$
end while

4. 実験

4.1 データセット

データセットには顔画像データセットとして知られるCelebA [17] を用いる．このデータセットは 20 万枚以上の顔画像データを含み，全ての画像に 40 種類の二値属性ラベルが付与されている．実験には約 16 万枚の画像を訓練データとして，残りの画像をテストデータとして用いる．このデータセットは顔方位変換画像生成モデルの学習に必要である顔方位情報 (yaw，pitch，roll) を持たないため，顔方位推定を行う推定モデル Hopenet [23] を用いて全ての顔画像にラベル付けする．方位ラベルの全ての顔方位角が 0 度であるとき，顔画像は正面を向いているとする．Hopenet の推定精度を考慮して，yaw 角の推定値が正面から ±45 度以内である顔画像を学習に使用する．

4.2 訓練

入力画像サイズは 96×96 のカラー画像でありデータセットから切り抜いたものを使用する．画像のピクセル値は $[-1, 1]$ に正規化する．データセットの方位ラベルは $[-1, 1]$ の範囲に正規化され，ラベルの値が 1 のとき各方位角は 45 度を表す．属性ラベルは 0 または 1 の値をとる 40 次元の二値ラベルとして使用する．ミニバッチサイズは 32 であり，全ての重みは平均 0，標準偏差 0.02 の正規分布によって初期化される．最適化アルゴリズムに Adam を用い，$\alpha = 10^{-4}$，$\beta_1 = 0.9$，$\beta_2 = 0.999$ とする．

4.3 生成結果

図 2 にランダムな潜在変数 z と様々な yaw 角の方位条件 y^P から G が生成した画像を示す．図 3 に CelebA に含まれる画像から E と G によって再構成された画像と，その再構成画像を入力として yaw 角の方位条件の変化毎に G が生成した画像を示す．これらの結果から提案モデルの生成画像は入力方位条件によらず個性を保存した結果を出力しており，特に入力画像に近い方位条件によって生成された再構成画像は入力画像をよく再現していることがわかる．全ての合成画像で目や鼻，口といった顔の各パーツについては入力方位条件を反映するが，髪型や顔の輪郭のような情報について入力方位条件を考慮して自然な画像として生成することは難しいことがわかる．入力画像が正面顔であるときは入力方位条件の変化による髪型や輪郭の生成は自然な画像として許容できるが，入力画像がやや横顔であり入力画像の方位と入力方位条件の符号が異なるときは，生成画像が大きく崩れてしまう傾向がある．

5. 結論と今後の課題

標準的な顔画像データセットを用いて顔方位を制御した画像生成を行う深層学習モデルの学習フレームワークを提案した．入力方位条件を考慮した画像生成のための条件損失とある潜在変数からの生成顔画像が入力方位条件によらず同一人物であるための属性損失を用いて，方位変換後の目的画像が存在しないデータセットによる学習から人物の個性表現と方位表現を分離した生成結果を得た．提案モデルは入力方位条件の変化によって人物の個性を変化させず，目や鼻，口などの顔の各パーツについて入力方位条件を反映した画像を生成するが，現状では入力方位条件を考慮した髪型や顔の輪郭の生成は困難であり，GAN による学習によって入力方位条件を考慮した自然な髪型や輪郭を生成するように改善することが課題として挙げられる．また今回の実験において生成画像の方位変化は ±45 度以内という制限があるが，大きな方位変換を行うことが可能な生成モデルへの改良も今後の課題である．

参考文献

[1] Bao, J., Chen, D., Wen, F., Li, H. and Hua, G.: CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training, *CoRR*, Vol. abs/1703.10155 (online), available from ⟨http://arxiv.org/abs/1703.10155⟩ (2017).

[2] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I. and Abbeel, P.: InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, *CoRR*, Vol. abs/1606.03657 (online), available from ⟨http://arxiv.org/abs/1606.03657⟩ (2016).

[3] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database, *CVPR09* (2009).

Vol.2018-CG-172 No.11
Vol.2018-DCC-20 No.11
Vol.2018-CVIM-214 No.11
2018/11/8

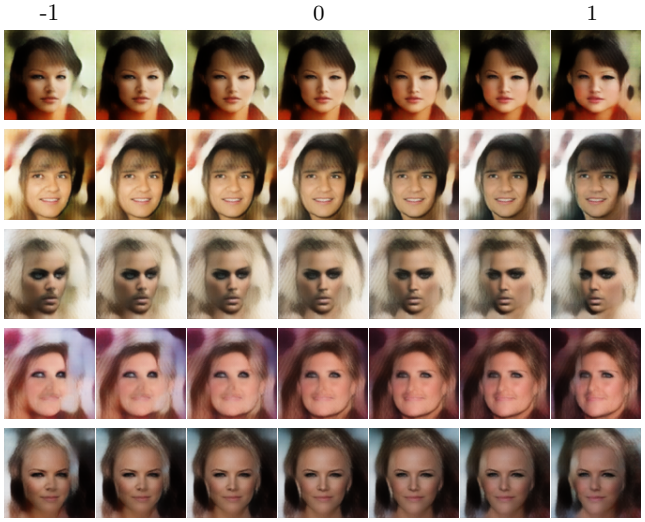

図 2　ランダムな潜在変数から生成された顔画像

Fig. 2　Syntheses from latent code sampled from prior distribution

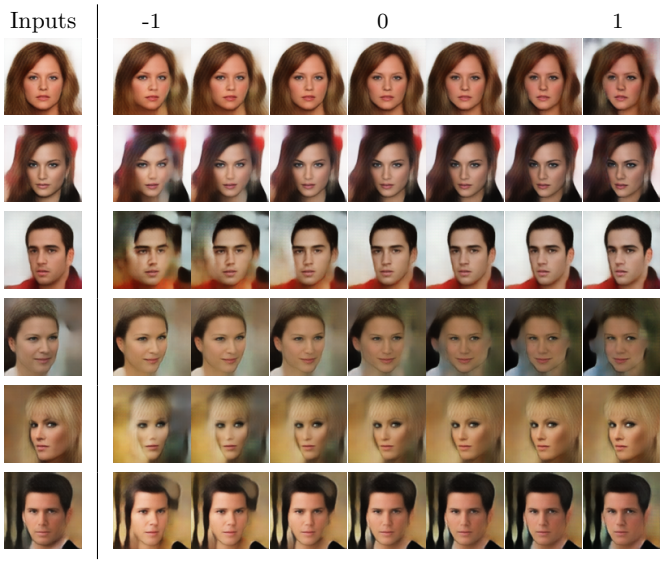

図 3　CelebA からの再構成画像を入力として方位条件の変化毎に生成された画像

Fig. 3　Reconstructions from CelebA and the syntheses with pose code

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative Adversarial Nets, *NIPS*, pp. 2672–2680 (online), available from ⟨http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf⟩ (2014).

[5] Gross, R., Matthews, I., Cohn, J., Kanade, T. and Baker, S.: Multi-PIE, *Image Vision Comput.*, Vol. 28, No. 5, pp. 807–813 (online), DOI: 10.1016/j.imavis.2009.08.002 (2010).

[6] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *CoRR*, Vol. abs/1512.03385 (online), available from ⟨http://arxiv.org/abs/1512.03385⟩ (2015).

[7] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework, *In Proceedings of the International Conference on Learning Representations (ICLR)* (2017).

[8] Hou, X., Shen, L., Sun, K. and Qiu, G.: Deep Feature Consistent Variational Autoencoder, *CoRR*, Vol. abs/1610.00291 (online), available from ⟨http://arxiv.org/abs/1610.00291⟩ (2016).

[9] Huang, G. B., Ramesh, M., Berg, T. and Learned-Miller, E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report 07-49, University of Massachusetts, Amherst (2007).

[10] Huang, R., Zhang, S., Li, T. and He, R.: Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis, *CoRR*, Vol. abs/1704.04086 (online), available from ⟨http://arxiv.org/abs/1704.04086⟩ (2017).

[11] Ioffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, *CoRR*, Vol. abs/1502.03167 (online), available from ⟨http://arxiv.org/abs/1502.03167⟩ (2015).

[12] Kan, M., Shan, S., Chang, H. and Chen, X.: Stacked Progressive Auto-Encoders (SPAE) for Face Recognition Across Poses, *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, Washington, DC, USA, IEEE Computer Society, pp. 1883–1890 (online), DOI: 10.1109/CVPR.2014.243 (2014).

[13] Kingma, D. P., Rezende, D. J., Mohamed, S. and Welling, M.: Semi-Supervised Learning with Deep Generative Models, *CoRR*, Vol. abs/1406.5298 (online), available from ⟨http://arxiv.org/abs/1406.5298⟩ (2014).

[14] Kingma, D. P. and Welling, M.: Auto-Encoding Variational Bayes, *CoRR*, Vol. abs/1312.6114 (online), available from ⟨http://arxiv.org/abs/1312.6114⟩ (2013).

[15] Larsen, A. B. L., Sønderby, S. K. and Winther, O.: Autoencoding beyond pixels using a learned similarity metric, *CoRR*, Vol. abs/1512.09300 (online), available from ⟨http://arxiv.org/abs/1512.09300⟩ (2015).

[16] Lin, M., Chen, Q. and Yan, S.: Network In Network, *CoRR*, Vol. abs/1312.4400 (online), available from ⟨http://arxiv.org/abs/1312.4400⟩ (2013).

[17] Liu, Z., Luo, P., Wang, X. and Tang, X.: Deep Learning Face Attributes in the Wild, *Proceedings of International Conference on Computer Vision (ICCV)* (2015).

[18] Makhzani, A., Shlens, J., Jaitly, N. and Goodfellow, I. J.: Adversarial Autoencoders, *CoRR*, Vol. abs/1511.05644 (online), available from ⟨http://arxiv.org/abs/1511.05644⟩ (2015).

[19] Mescheder, L. M., Nowozin, S. and Geiger, A.: Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks, *CoRR*, Vol. abs/1701.04722 (online), available from ⟨http://arxiv.org/abs/1701.04722⟩ (2017).

[20] Mirza, M. and Osindero, S.: Conditional Generative Adversarial Nets, *CoRR*, Vol. abs/1411.1784 (online), available from ⟨http://arxiv.org/abs/1411.1784⟩ (2014).

[21] Miyato, T., Kataoka, T., Koyama, M. and Yoshida, Y.: Spectral Normalization for Generative Adversarial Networks, *CoRR*, Vol. abs/1802.05957 (online), available from ⟨http://arxiv.org/abs/1802.05957⟩ (2018).

[22] Radford, A., Metz, L. and Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,

CoRR, Vol. abs/1511.06434 (online), available from ⟨http://arxiv.org/abs/1511.06434⟩ (2015).

[23] Ruiz, N., Chong, E. and Rehg, J. M.: Fine-Grained Head Pose Estimation Without Keypoints, *CoRR*, Vol. abs/1710.00925 (online), available from ⟨http://arxiv.org/abs/1710.00925⟩ (2017).

[24] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, Vol. abs/1409.1556 (online), available from ⟨http://arxiv.org/abs/1409.1556⟩ (2014).

[25] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: Rethinking the Inception Architecture for Computer Vision, *CoRR*, Vol. abs/1512.00567 (online), available from ⟨http://arxiv.org/abs/1512.00567⟩ (2015).

[26] Tran, L., Yin, X. and Liu, X.: Disentangled Representation Learning GAN for Pose-Invariant Face Recognition, *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Honolulu, HI (2017).

[27] Tran, L., Yin, X. and Liu, X.: Representation Learning by Rotating Your Faces (2018).

[28] Yang, J., Reed, S. E., Yang, M. and Lee, H.: Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis, *CoRR*, Vol. abs/1601.00706 (2016).

[29] Yim, J., Jung, H., Yoo, B., Choi, C., Park, D.-S. and Kim, J.: Rotating your face using multi-task deep neural network., *CVPR*, IEEE Computer Society, pp. 676–684 (2015).

[30] Yin, X., Yu, X., Sohn, K., Liu, X. and Chandraker, M.: Towards Large-Pose Face Frontalization in the Wild, *CoRR*, Vol. abs/1704.06244 (2017).

[31] Zheng, Z., Yu, Z., Zheng, H., Wang, C. and Wang, N.: Pipeline Generative Adversarial Networks for Facial Images Generation with Multiple Attributes, *CoRR*, Vol. abs/1711.10742 (online), available from ⟨http://arxiv.org/abs/1711.10742⟩ (2017).

[32] Zhu, Z., Luo, P., Wang, X. and Tang, X.: Deep Learning Multi-View Representation for Face Recognition, *CoRR*, Vol. abs/1406.6947 (2014).