2ZD-01

収集コストの低いデータセットを用いた顔方位変換画像生成モデル

森川将平 † 齋藤豪 †
† 東京工業大学 情報理工学院

1 はじめに

深層学習を用いた顔方位変換画像の生成モデルでは、人物を多視点から撮影することによって作成された Multi-PIE [1] に代表される顔画像データセットを用いることが一般的である。従来手法では顔方位変換を行った生成画像と目的画像の差分や、方位変換前後の顔画像間における同一人物性の判定に個人識別用の ID ラベルを用いてネットワークの学習を行う。しかし、人物の顔画像に対してその方位変換後の目的画像を用意することや、個人識別用のラベルを各顔画像に付与することは一般的に容易ではなく、従来手法で用いられるデータセットはデータの収集や追加拡張が困難であるといえる。

そこで本稿ではそのようなデータを用いずに顔方位変換によって同一人物性を失わないための損失関数を定義し、収集コストの低いデータのみを用いて、入力顔画像に対して顔方位変換を行う画像生成モデルを提案する。

2 提案手法

2.1 データセット

本稿では CelebA データセット [2] と YouTubeFaces データベース [3] を用いて学習を行う。CelebA は有名人の顔画像を収集した顔画像データセットであり、YouTubeFaces は YouTube から収集した顔動画データセットである。どちらのデータセットにも動画像の人物を特定する ID ラベルが付与されている。しかしこれらのラベルデータは画像や動画の収集とは異なり、新規のデータ作成が簡単でないデータであるとして本手法では用いない。顔方位変換画像の生成において顔画像に対する顔方位情報が必要であるが、本稿では入力された顔画像の顔方位を推定する深層学習モデルである Hopenet [4] を用いて CelebA データセットの各顔画像に対して自動アノテーションを行う。提案モデルはこれらのデータセットの動画像と用意した方位ラベルのみを用いて訓練される。

2.2 ネットワークアーキテクチャ

本稿で提案するモデルは生成モデルである VAE と GAN を基本構造に持ち、ネットワーク概要（図 1）に示されるように、Encoder，Generator，Predictor，VGG16net3 つの Discriminator の計 7 つのサブネットワークから構成される。それぞれのネットワークは E，G，P，VGG，D_{Code}，D_{Img}，D_{ID} で表される。

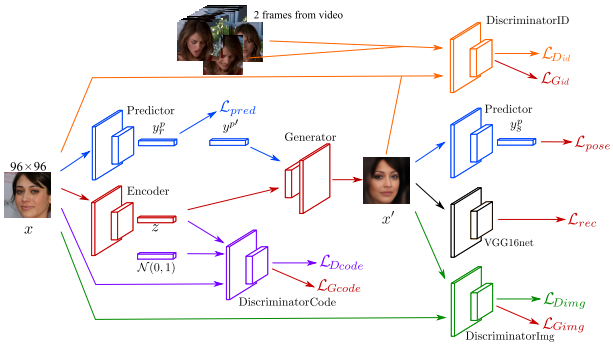

図 1: ネットワーク概要

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 [5] で提案される VGG 損失を用いる。

$$L_{rec} = \sum_n ||\Phi_n(x) - \Phi_n(G(z_x, y_x^p))||_1 \quad (1)$$

ここで $\Phi_n(x)$ は画像 x を VGG16 へ入力した際の中間層 Block n-Conv1 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。

本手法では 3 つの Discriminator を用いて敵対的学習を行っている。$L_{D\_}$ はそれぞれの Discriminator を訓練するための損失関数であり、$L_{G\_}$ は訓練された Discrminator を騙すようなデータの生成を行う Encoder や Generator のための損失関数である。

D_{code} は Adversarial Variational Bayes [6] で提案される Discriminator であり、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。また E は D_{code} を騙すような潜在変数を生成するように学習し、潜在空間の確率分布をモデル化する。

$$\begin{aligned} L_{D_{Code}} = &- \mathbb{E}_{z\sim p_z(z)}[\log D_{code}(z)] \\ &- \mathbb{E}_{x\sim p_{data}(x)}[\log(1 - D_{code}(z_x))] \quad (2) \\ L_{G_{code}} = &- \mathbb{E}_{x\sim p_{data}(x)}[\log D_{code}(z_x)] \quad (3) \end{aligned}$$

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。

Generative model for face rotation images with dataset collected with less effort
† Shohei MORIKAWA
† Suguru SAITO
School of Computing, Tokyo Institute of Technology (†)

D_{Img} は従来の GAN で提案される Discriminator であり、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する。また G は D_{Img} を騙すようにデータセットと見分けが付かないような合成画像を生成するように学習する。

$$L_{D_{Img}} = -\mathbb{E}_{x\sim p_{data}(x)}[\log D_{Img}(x)] - \mathbb{E}_{z\sim p_z(z)}[\log(1 - D_{Img}(G(z, y^p)))] \quad (4)$$

$$L_{G_{Img}} = -\mathbb{E}_{z\sim p_z(z)}[\log D_{Img}(G(z, y^p))] \quad (5)$$

D_{ID} は本手法で提案する方位変換前後の同一人物性を計るための Discriminator であり、入力された 2 枚の顔画像が同一人物であるかどうかを二値判定する。ここでは動画が連続した静止画の集合であるため、あるシーンから切り抜かれた 2 枚の顔画像が同一人物であるという性質を活かしネットワークを訓練する。従来手法では ID ラベルによる分類によって生成画像が正しく個人識別されるように学習していたが、本手法では G が D_{ID} を騙すように入力画像の自然な方位変換によって得られる顔画像を生成するように学習する。

$$L_{D_{ID}} = -\mathbb{E}_{x\sim p_{data}(x)}[\log(1 - D_{ID}(x, G(z_x, y^p)))] - \mathbb{E}_{x\sim p_{video}(x)}[\log(1 - D_{ID}(x_i^k, x_j^l))] - \mathbb{E}_{x\sim p_{video}(x)}[\log D_{ID}(x_i^k, x_i^l)] \quad (6)$$

$$L_{G_{ID}} = -\mathbb{E}_{x\sim p_{data}(x)}[\log D_{ID}(x, G(z_x, y^p))] \quad (7)$$

ここで $x_n^m \sim p_{video}(x)$ は YouTubeFaces からサンプリングされた、動画のあるシーン n における m 番目のフレームから顔領域を切り抜いて得られる顔画像である。

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を推定する。P は G の学習のために後述される条件損失 L_{pose} を算出するために L_{pred} によって訓練される。L_{pred} はラベル付けされた方位情報と生成画像の方位推定値の $L2$ ノルムを用いて定義される。

$$L_{pred} = ||y_x^p - P_{pose}(x)||_2 \quad (8)$$

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} を考える。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。

$$L_{pose} = ||y^p - P_{pose}(G(z, y^p))||_2 \quad (9)$$

3 結果

図 2 に入力顔画像と方位条件の変化毎に G が生成した画像を示す。上下段は本稿で提案した DiscriminatorID を用いた損失関数の有無によって学習した 2 つのモデルの生成結果である。上下段ともに入力方位条件に沿った顔方位を持つ画像を生成しており、入力画像に近い方位条件による生成画像は入力画像をよく表現していることがわかる。しかし式 (6)(7) で定義される損失関数を用いない下段の結果では入力方位条件の変化によって入力画像の個性が保存されていない。一方で上段の結果では下段に比べて入力方位条件の変化によらず入力画像の個性を保存した方位変換画像が生成されていることがわかる。

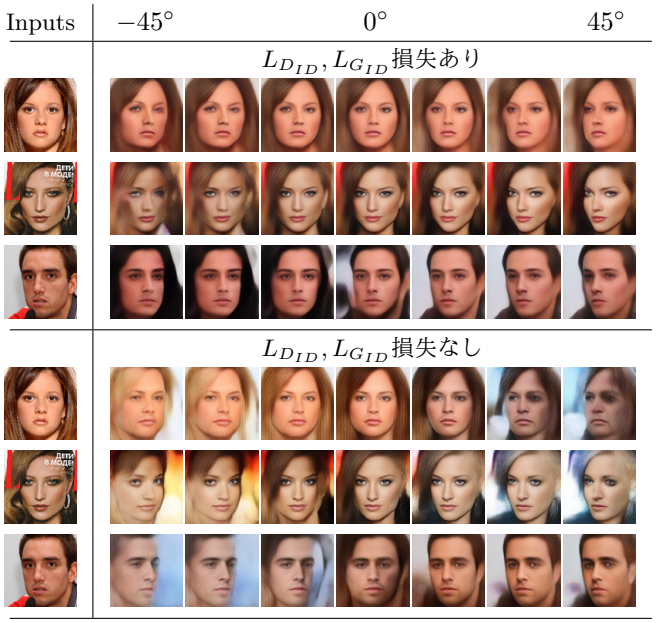

図 2: 入力顔画像と方位条件毎に生成された画像

4 まとめ

顔方位変換画像生成モデルにおいて顔方位変換後の目的画像や各顔画像に対する ID ラベルのような収集コストの高いデータを学習に必要としていた従来法に対して、顔動画データのような収集コストの低いデータのみ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。提案した損失関数を用いずに学習したモデルに比べて、提案モデルでは人物の個性表現と方位表現を分離した生成結果が得られた。しかし入力画像に対する方位の変化量が大きくなるほど生成画像は顔の輪郭や髪型の保持が困難であるため、敵対的学習の改善によって入力方位条件を考慮した、より自然な画像を生成することが今後の課題である。

参考文献

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

[2] Liu, Z., Luo, P., Wang, X. and Tang, X.: Deep Learning Face Attributes in the Wild, *ICCV* (2015).

[3] Wolf, L., Hassner, T. and Maoz, I.: Face recognition in unconstrained videos with matched background similarity, pp. 529–534 *CVPR*,(2011).

[4] Ruiz, N., Chong, E. and Rehg, J. M.: Fine-Grained Head Pose Estimation Without Keypoints, *CVPR* (2018).

[5] Hou, X., Shen, L., Sun, K. and Qiu, G.: Deep Feature Consistent Variational Autoencoder, *CoRR*, Vol. abs/1610.00291 (online), (2016).

[6] Mescheder, L. M., Nowozin, S. and Geiger, A.: Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks, *CoRR*, Vol. abs/1701.04722 (online), (2017).