Vol.2018-CVIM-212 No.28
2018/5/10

RGB-D動画生成のためのGAN

中平有樹[1,a)]　川本一彦[2,b)]

概要：深層生成モデルである GAN（Generative Adversarial Networks）は，これまで画像，動画，そして 3D ボリュームデータの生成に適用されてきたが，RGB-D 動画ではほとんど例がない．本研究では，RGB-D 動画生成のための 4 つの GAN モデルを比較検証する．4 つのモデルは，（1）最も基本的な GAN の畳み込み演算を 3 次元に拡張したモデル，（2）前景と背景を分けて生成する VideoGAN に深度チャンネルを付加したモデル，（3）動きと内容に分けて潜在空間に埋め込み RNN で時系列モデリングする MoCoGAN に深度チャンネルを付加したモデル，（4） MoCoGAN をベースにしつつ，RGB と depth をそれぞれ独立に生成するように拡張したモデル，である．顔表情変化の動画データセットに対して，主観評価と Inception Score を用いた客観評価で比較し，RGB と depth を独立に扱わない MoCoGAN に基づくモデルで最も高品質な動画を生成できることが判明した．

キーワード：GAN，動画生成，深層学習，RGB-D データ

1. はじめに

コンピュータビジョンにおいて RGB-D データは広く用いられており，それらのデータを自動的に生成する試みは非常に重要である．加えて近年では深層学習が盛んに用いられており，大量の学習データが必要となるため，データ生成の必要性はますます高まっているといえる．

GAN（Generative Adversarial Nets）[1] は，ニューラルネットワークによってデータの確率分布の推定を行う深層生成モデルの一つである．大量のデータを学習し，そのデータと同じような新しいデータを生成出来ることが大きな特徴である．これまで，GAN を用いた静止画生成に関する研究が盛んに行われてきた．例えば白黒画像に着色を行う pix2pix[2]，シマウマの画像を馬の画像に変換する CycleGAN[3]，画像の超解像を行う SRGAN[4]，海外の有名人の高解像度 (1024 × 1024) な顔画像を生成する Progressive GAN[5] などである．これらの研究より，GAN による静止画の生成は非常に高精度に行えるようになってきていると言える．

GAN によるデータ生成は静止画にとどまらず，動画の生成にも拡張され始めている．Vondrick らは，ビデオ生成に GAN を初めて適用し，動画を前景と背景に分けてモデル化する VideoGAN[6] を提案した．また Tulyakov らは，動画を内容（Content）と動き（Motion）に分けてモデル化する MoCoGAN[7] を提案している．

しかし，これまでに RGB-D 動画の生成に GAN を適用した研究は存在しない．よって本稿では，通常の動画生成を基に RGB-D 動画を生成するための GAN モデルを 4 つ提案し，比較検証する．4 つのモデルは，（1）最も基本的な GAN の畳み込み演算を 3 次元に拡張したモデル，（2）前景と背景を分けて生成する VideoGAN に深度チャンネルを付加したモデル，（3）動きと内容に分けて潜在空間に埋め込み RNN で時系列モデリングする MoCoGAN に深度チャンネルを付加したモデル，（4） MoCoGAN をベースにしつつ，RGB と depth をそれぞれ独立に生成するように拡張したモデル，である．

2. GANによる動画生成

2.1 GAN

GAN は generator と discriminator の 2 つのネットワークで構成されるモデルである．アーキテクチャを**図 1** に示す．generator は

$$\tilde{\boldsymbol{x}} = G(\boldsymbol{z}) \tag{1}$$

のように，G を generator とした時，任意の分布からのサンプル $\boldsymbol{z}$ を入力として受け取り，データ $\boldsymbol{x}$ を生成する．これに対し，discriminator はデータを入力として受け取り，

[1] 千葉大学大学院 融合理工学府
Graduate School of Science and Engineering, Chiba University
[2] 千葉大学大学院 工学研究院
Graduate School of Engineering, Chiba University
[a)] nakahira@chiba-u.jp
[b)] kawa@faculty.chiba-u.jp

ⓒ 2018 Information Processing Society of Japan

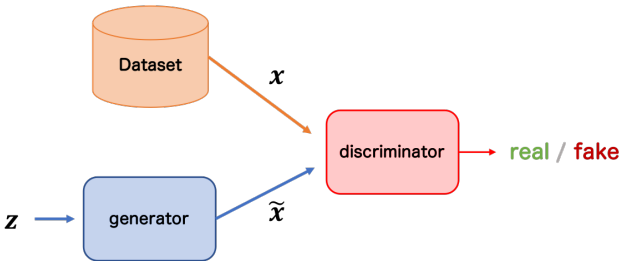


図 1: GAN のアーキテクチャ

それが学習データセットのサンプル $\boldsymbol{x}$ であるか，generator によって生成されたデータ $\tilde{\boldsymbol{x}}$ かの 2 値分類を行う．GAN はこの 2 つのネットワークを互いに競わせることで学習を行う．これは

$$\min_G \max_D V(D,G), \tag{2}$$
$$V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_g(\boldsymbol{x})}[\log(1 - D(\tilde{\boldsymbol{x}}))]$$

のように表される．ここで，D は discriminator，p_g は generator の分布，p_{data} は学習データの分布を示す．

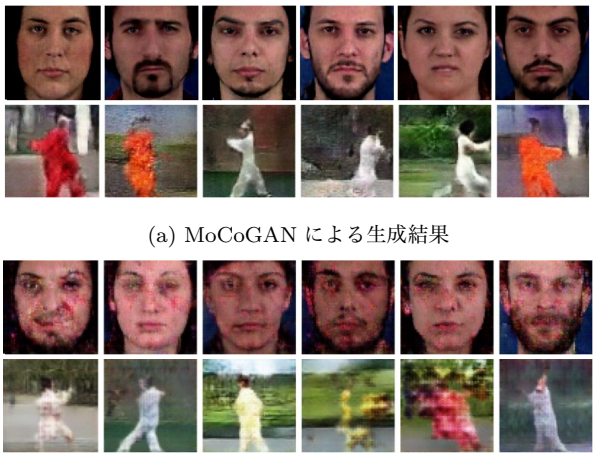
2.2 VideoGAN

2016 年，Vondrick らによってビデオ生成に GAN を初めて適用した VideoGAN[6] が提案された．動画生成において重要となるのは，GAN でダイナミクスのモデリングをどのように行うかということである．VideoGAN では，固定されたカメラで撮影した動画が，動的な前景（Foreground）と静的な背景（Backgroud）に分けられることを指摘し，この知識を反映させたアーキテクチャを提案している．具体的には，前景にあたる動画は 3 次元転置畳み込み，背景にあたる静止画は 2 次元転置畳み込みを用いてそれぞれ生成し，最後に前景と背景を足し合わせて動画を出力する．学習では，前景と背景を足し合わせる際の重みもアーキテクチャに組み込んで同時に学習を行う．discriminator は通常の GAN と同様であり，入力された動画がデータセット由来か，generator 由来かを判別する役割を担っている．

VideoGAN は，前景と背景を分けるアーキテクチャにより，多くのシーンで背景は鮮明なものが生成され，動画内の適切な物体にシーンに適した動きも付加できる．一方，生成される動きのある物体の解像度が低いことが問題であり，例えば，人間が人間らしく生成できていないサンプルも生成される．このことから前景の生成に課題があるとされている．

2.3 MoCoGAN

2017 年に，Tulyakov らによって MoCoGAN が提案された [7]．Tulyakov らは，VideoGAN のような潜在空間上の 1 点が 1 動画に対応するモデルは，同じ動きを異なる速度で実施した動画がそれぞれ潜在空間上で別の点となってしまう問題を指摘した．これを改善するために，MoCoGAN では動画が内容（Content）と動き（Motion）の 2 つに分解できることを指摘し，この知識を反映させたアーキテクチャを提案している．具体的には，動画の内容に対応するノイズベクトルと動きに対応するノイズベクトルの 2 つをそれぞれ generator に入力することで動画を生成する．これにより，例えば人の表情の動画生成においては，顔の個人性（内容）を固定して，異なる表情（動き）を生成するといった制御が容易になる．



(a) MoCoGAN による生成結果

(b) VideoGAN による生成結果

図 2: MoCoGAN の生成結果（[7] より転載）

MoCoGAN と VideoGAN の生成結果の比較を**図 2** に示す．VideoGAN はシーン動画の生成を想定したアーキテクチャであるため，顔表情動画の生成は不向きである可能性もあるが，VideoGAN に比べると MoCoGAN はよりきれいな動画を生成する傾向があることがわかる [7]．クラウドソーシングを用いた主観評価実験においても VideoGAN を上回ることが示されている．

3. 提案手法

本節では，前節で述べた基本的な GAN，VideoGAN，MoCoGAN を基にし，新たに RGB-D 動画を生成するための GAN アーキテクチャを 4 つ提案する．

3.1 RGBD-GAN

GAN の最も基本的な形は generator と discriminator がそれぞれ一つずつあるようなモデルである．これを RGB-D 動画の生成のために畳み込み演算を 3 次元に拡張し，深度チャネルを付加したモデルを提案する．アーキテクチャを**図 3a** に示す．図中の 3D は空間方向と時間方向を合わせた 3 次元の畳込みを内部で行うことを示す．以後これを RGBD-GAN と呼ぶ．

RGBD-GAN モデルにおいてサンプルは

$$\tilde{x} = G(\boldsymbol{z}) \tag{3}$$

のように生成される．ここで，G は generator である．学習は

$$\min_G \max_D V(D,G), \tag{4}$$
$$V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_g(\boldsymbol{x})}[\log(1-D(\tilde{\boldsymbol{x}}))]$$

を最適化することによって行われる．ここで，D は discriminator，p_g は generator の分布，p_{data} は学習データの分布を示す．

3.2 RGBD-VideoGAN

これは 2.2 節で述べた VideoGAN のアーキテクチャを RGB-D 動画の生成へと拡張したモデルである．このアーキテクチャを**図 3b** に示す．図中の 2D は空間方向のみの 2 次元畳み込み，3D は空間方向と時間方向を合わせた 3 次元の畳み込みを内部で行うことを示す．以後これを RGBD-VideoGAN と呼ぶ．

RGBD-VideoGAN モデルにおいてサンプルは

$$\tilde{\boldsymbol{x}} = m(\boldsymbol{z}) \odot G_f(\boldsymbol{z}) + (1-m(\boldsymbol{z})) \odot G_b(\boldsymbol{z}). \tag{5}$$

のように生成される．ここで，G_f は動画の前景を生成する generator，G_b は動画の背景を生成する generator であり，m は G_f の最終層で分岐し前景と背景を足し合わせる割合を推定するネットワークである．また $\odot$ は要素同士の掛け算を示す．学習は式 (4) を最適化することで行われる．

3.3 RGBD-MoCoGAN

これは 2.3 節で述べた MoCoGAN のアーキテクチャを RGB-D 動画の生成へと拡張したモデルである．このアーキテクチャを**図 3c** に示す．図中の 2D は空間方向のみの 2 次元畳み込み，3D は空間方向と時間方向を合わせた 3 次元の畳み込みを内部で行うことを示す．以後これを RGBD-MoCoGAN と呼ぶ．

RGBD-MoCoGAN モデルにおいて，サンプルは

$$\tilde{\boldsymbol{x}} = \left[G_I(\begin{bmatrix}\boldsymbol{z_C}\\ \boldsymbol{z_M^{(1)}}\end{bmatrix}), ..., G_I(\begin{bmatrix}\boldsymbol{z_C}\\ \boldsymbol{z_M^{(T)}}\end{bmatrix}) \right] \tag{6}$$

のように生成される．ここで，

$$\left[\boldsymbol{z_M^{(1)}}, ..., \boldsymbol{z_M^{(T)}}\right] = \left[R_M(\boldsymbol{h_0}, \boldsymbol{\epsilon^{(1)}}), ..., R_M(\boldsymbol{z_M^{(T-1)}}, \boldsymbol{\epsilon^{(T)}}) \right]$$

であり，R_M は動画の動きに相当するベクトルを生成するネットワークを，G_I は動画のフレームを生成する generator である．T は動画長を示す．学習は

$$\min_{G_I, R_M} \max_{D_I, D_V} V(G_I, R_M, D_I, D_V), \tag{7}$$
$$V(G_I, R_M, D_I, D_V) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D_I(\boldsymbol{x}^{(t)})] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_g}[\log(1-D_I(\tilde{\boldsymbol{x}}^{(t)}))] + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D_V(\boldsymbol{x})] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_g}[\log(1-D_V(\tilde{\boldsymbol{x}}))]$$

を最適化することで行われる．ここで，D_I は image discriminator，D_V は video discriminator，p_g は generator の分布，p_{data} は学習データの分布を示す．

3.4 RGB+D-MoCoGAN

前節までの 3 つのモデルでは RGB と depth を同一の generator で生成していた．しかし RGB は色であり，depth は奥行き（距離）であることから，これらは本質的に異なる情報であると考えられる．よって，MoCoGAN のアーキテクチャを基本とし，RGB を生成する generator と depth を生成する generator をそれぞれ独立に持つアーキテクチャを提案する．これを**図 3d** に示す．図中の 2D は空間方向のみの 2 次元畳み込み，3D は空間方向と時間方向を合わせた 3 次元の畳み込みを内部で行うことを示す．以後これを RGB+D-MoCoGAN と呼ぶ．

RGB+D-MoCoGAN モデルにおいて，サンプルは

$$\tilde{\boldsymbol{x}} = \left[\tilde{\boldsymbol{x}}^{(1)}_{\text{rgb}} \oplus \tilde{\boldsymbol{x}}^{(1)}_{\text{depth}}, ..., \tilde{\boldsymbol{x}}^{(T)}_{\text{rgb}} \oplus \tilde{\boldsymbol{x}}^{(T)}_{\text{depth}} \right] \tag{8}$$
$$\tilde{\boldsymbol{x}}_{\text{rgb}} = \left[G_{\text{rgb}}(\begin{bmatrix}\boldsymbol{z_C}\\ \boldsymbol{z_M^{(1)}}\end{bmatrix}), ..., G_{\text{rgb}}(\begin{bmatrix}\boldsymbol{z_C}\\ \boldsymbol{z_M^{(T)}}\end{bmatrix}) \right]$$
$$\tilde{\boldsymbol{x}}_{\text{depth}} = \left[G_{\text{depth}}(\begin{bmatrix}\boldsymbol{z_C}\\ \boldsymbol{z_M^{(1)}}\end{bmatrix}), ..., G_{\text{depth}}(\begin{bmatrix}\boldsymbol{z_C}\\ \boldsymbol{z_M^{(T)}}\end{bmatrix}) \right]$$
$$\left[\boldsymbol{z_M^{(1)}}, ..., \boldsymbol{z_M^{(T)}}\right] = \left[R_M(\boldsymbol{h_0}, \boldsymbol{\epsilon^{(1)}}), ..., R_M(\boldsymbol{z_M^{(T-1)}}, \boldsymbol{\epsilon^{(T)}}) \right]$$

のように生成される．ここで，R_M は動画の Motion に相当するベクトルを生成するネットワークを，G_{rgb} は動画の各フレームの RGB チャネルを生成する generator，G_{depth} は depth チャネルを生成する generator である．$\oplus$ はチャネル方向の結合（concatenation），T は動画長を示す．学習は

$$\min_{G_{\text{rgb}}, G_{\text{depth}}, R_M} \max_{D_I, D_V} V(G_{\text{rgb}}, G_{\text{depth}}, R_M, D_I, D_V), \tag{9}$$
$$V(G_{\text{rgb}}, G_{\text{depth}}, R_M, D_I, D_V) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D_I(\boldsymbol{x}^{(t)})] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_g}[\log(1-D_I(\tilde{\boldsymbol{x}}^{(t)}))] + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D_V(\boldsymbol{x})] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_g}[\log(1-D_V(\tilde{\boldsymbol{x}}))]$$

を最適化することで行われる．ここで，D_I は image discriminator，D_V は video discriminator，p_g は generator の

分布，p_{data} は学習データの分布を示す．

4. 評価実験

本節では，前節で提案した 4 つの GAN アーキテクチャを人の顔表情変化の動画データセットを用いて訓練し，生成結果の比較を行う．比較の際には実際の生成結果を示した上で，著者自身による主観評価と Inception Score を用いた客観評価を行う．

4.1 データセット

本実験では，既存の RGB 動画データセットに対して 3 次元復元を適用し，予測された 3 次元情報から深度情報を得ることで，RGB-D 動画データセットを新たに作成した．基となるデータセットは MUG Facial Expression Dataset[8] であり，これは男女 86 人（男性 51 人，女性 35 人）が様々な表情を演じるデータセットである．それぞれの動画は画像サイズ 896 × 896，ビデオ長 50〜160 フレームで構成されており，表情は怒り，嫌悪感，恐怖，喜び，悲しみ，驚きの 6 種類が収録されている．これに対し，以下の前処理を各フレームに対して行うことで深度情報を算出する．

(1) 顔検出を行い，顔の領域のみをクロッピングする
(2) CNN を用いた 3 次元復元手法 [9] を適用し，$(height, width, depth) = (192, 192, 200)$ のようなボクセルを得る
(3) ボクセルの奥行き方向の最大値を深度とする．
(4) 1 で得られた RGB 画像と 2，3 で得られた深度画像をそれぞれ 64 × 64 にリサイズする．

最終的に，本データセットはサイズ 64 × 64 の RGB-D 動画データセットとなる．また，本実験では生成するビデオの長さが 16 で固定であるため，各動画から長さ 16 のサブシーケンスを複数抜き出して用いた．図 4 にデータセットサンプルの一例を示す．

4.2 実験条件

前節で述べた 4 つの GAN（RGBD-GAN，RGBD-VideoGAN, RGBD-MoCoGAN, RGB+D-MoCoGAN）でデータセットを学習し，生成結果の品質を比較する．

VideoGAN，MoCoGAN の元論文の構成にならい，全てのモデルで generator と discriminator は 5 層の転置畳み込み層・畳み込み層で構成される CNN とした．generator では ReLU と Batch Normalization を，discriminator では Leaky-ReLU と Batch Normalization を用いている．学習では，disciminator の学習が generator に比べて早く進んでしまうことから，安定化のために discriminator の各層に $\sigma = 0.2$ のガウスノイズを付加した [10]．ランダムベクトル $\boldsymbol{z}$ をサンプルする分布においても元論文を参考にし，全てのモデルにおいて平均 0，分散 0.33 のガウス分布を用い，RGBD-GAN モデルと RGBD-VideoGAN モデルでは 100 次元，RGBD-MoCoGAN モデルと RGB+D-MoCoGAN モデルでは $\boldsymbol{z_C}$ を 50 次元，$\boldsymbol{z_M}$ を 10 次元とした．最後にモデルの学習条件を表 1 に示す．

表 1: 実験条件

データセット	MUG Facial Expression Database 形式: RGB-D 動画 サンプル数:3219 (channel, frame, height, width) : (4, 16, 64, 64)
バッチサイズ	35
学習回数	200 epoch
最適化手法	Adam($\alpha = 0.0002, \beta_1 = 0.5, \beta_2 = 0.999$)

4.3 主観評価

学習された各モデルの generator で 100 サンプル生成し，生成結果の比較を行った．各モデルの生成結果の一部を**図 5** に示す．図では，出力サンプルが縦方向に RGB，depth のペアで 3 つ並んでおり，横方向には順に 4，8，12，16 番目のフレームが並んでいる．

これに対し，著者自身が目視で観察を行い品質の判定を行った．観察においては，RGB 動画の画質，depth による顔の形状，ダイナミクスの大きさの 3 点に着目した．この 3 点において，もっとも良い生成結果を示したのは RGBD-MoCoGAN モデルであった．RGBD-GAN モデルと RGBD-VideoGAN モデルは他に比べて明らかに RGB 動画の画質が低いことがわかった．RGB+D-MoCoGAN モデルは，RGBD-MoCoGAN モデルと同程度に画質は良かったものの，顔の形状（深度画像の品質）が悪いことがわかった．この主観評価の結果を簡潔に三段階（◯，△，×）で表したものを**表 2** に示す．

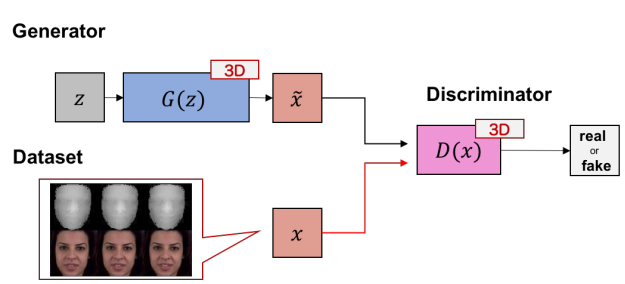
4.4 客観評価

本実験では GAN の生成結果の客観評価指標として Inception Score[11] を用いる．Inception Score は生成画像の品質を測る指標の 1 つであり，人による主観評価に代わるものとして Salimans らによって提案された．これは 2 つの確率分布の KL ダイバージェンスを求めたもので，式 (10) のように表される．

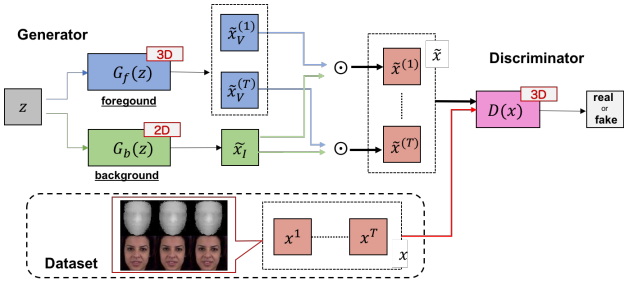
$$\exp(\mathbb{E}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}[\mathrm{KL}(p(y|\boldsymbol{x})||p(y)]) \qquad (10)$$

ここで，$\boldsymbol{x}$ は GAN による生成サンプルを，y はその所属クラスを示し，本実験では幸せや怒りなどの 6 種の表情となる．$p(y)$，$p(y|\boldsymbol{x})$ は事前に訓練された分類器によって算出される．この指標は，GAN によって生成された良い画像とは以下の 2 つ性質を持つと仮定している．

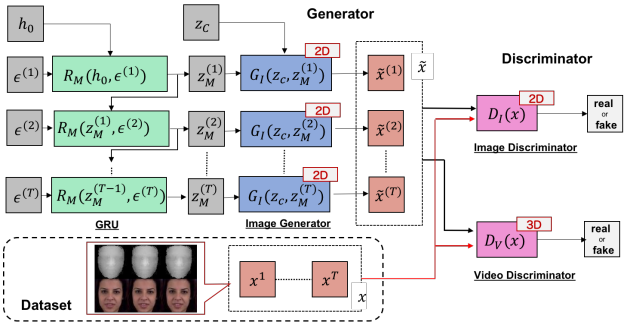
ⓒ 2018 Information Processing Society of Japan



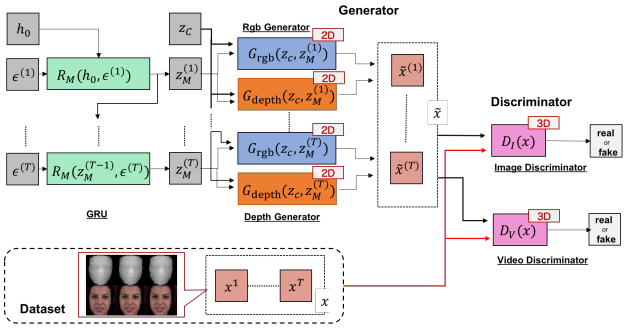
(a) RGBD-GAN のアーキテクチャ



(b) RGBD-VideoGAN のアーキテクチャ



(c) RGBD-MoCoGAN のアーキテクチャ



(d) RGB+D-MoCoGAN のアーキテクチャ

図 3: 各モデルのアーキテクチャ

表 2: 主観評価

	RGBD-GAN	RGBD-VideoGAN	RGBD-MoCoGAN	RGB+D-MoCoGAN
RGB 画質	×	×	○	○
depth による顔形状	○	×	○	×
ダイナミクスの大きさ	△	×	○	○

表 3: Inception Score による評価

Model	Inception Score	Rank
RGBD-GAN	4.73 ± 0.0149	2
RGBD-VideoGAN	4.55 ± 0.0157	4
RGBD-MoCoGAN	4.77 ± 0.00914	1
RGB+D-MoCoGAN	4.67 ± 0.0184	3
Dataset	5.17 ± 0.0351	-

(1) 高い信念でクラス分類ができること．すなわち，エントロピー $p(y|\boldsymbol{x})$ が小さいこと．
(2) サンプルのバリエーションが大きいこと．すなわち，エントロピー $p(y)$ が大きいこと．

算出に必要な分類器には，本実験と同様のデータセットで訓練された C3D[12] を用いた．GAN で 10000 サンプルを生成し，Inception Score を算出する手順を 4 回繰り返し，その平均と標準偏差を記録した．この結果を表 3 に示す．表 3 より，生成結果の品質は RGBD-MoCoGAN，RGBD-GAN，RGB+D-MoCoGAN，RGBD-VideoGAN の順に良いことがわかる．

4.5 考察

主観評価，客観評価の双方において，RGBD-MoCoGAN モデルによって生成された動画の品質が最も高いことがわかった．これは，そもそも MoCoGAN が基本的な GAN や VideoGAN に比べ，ダイナミクスのモデリングを適切に行えるアーキテクチャであるためと考えられる．今回の結果から，通常の動画生成を高品質に行えるモデルは，同様に RGB-D 動画の生成においても高い精度を発揮できると考えられる．

RGB+D-MoCoGAN は，MoCoGAN のアーキテクチャを踏襲しつつ，RGB と depth を生成する generator を分けることでより RGB-D 動画生成に特化することを狙いとしていたが，主観評価・客観評価ともに RGBD-MoCoGAN モデルよりも悪い結果となった．これは，RGB 画像と depth 画像には強い相関があるため，それらを独立にモデル化することによって逆に学習を難しくしてしまったことが原因と考えられる．

RGBD-GAN モデルは，主観評価では動画の画質が RGBD-MoCoGAN モデルや RGB+D-MoCoGAN モデルと比べ低いことが判明したが，一方で客観評価では RGBD-MoCoGAN モデルにせまる高い評価となった．これは，

ⓒ 2018 Information Processing Society of Japan



(a) anger



(b) disgust



(c) happiness



(d) fear



(e) sadness



(f) surprise

図 4: データセットのサンプル（上段は RGB 画像，下段は depth 画像）

Inception Score の算出に用いた分類器の性質に起因するものであると考えれる．用いた分類器は，与えられた動画の表情ラベルを推定するモデルであるが，表情の違いは主にダイナミクスの変化から判断できるため，多少画質が荒かった場合にも問題なく分類できる可能性がある．その場合，細かい画質の差が分類結果に影響しないため，Inception Score に正確に反映されなかったと考えられる．

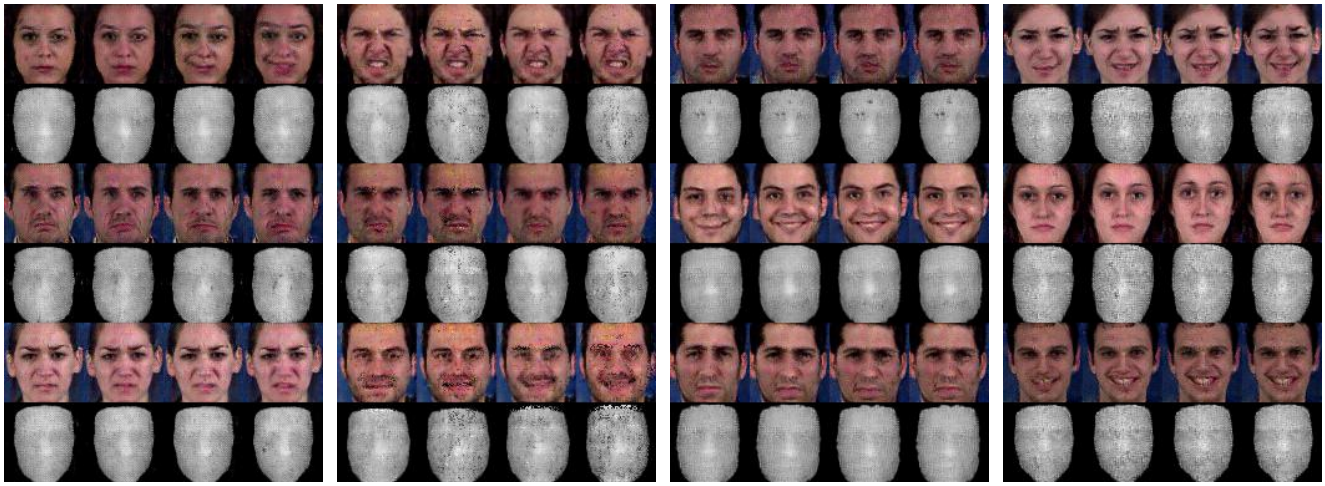
5. おわりに

本研究では，RGB-D 動画生成のための GAN アーキテクチャとして RGBD-GAN，RGBD-VideoGAN，RGBD-MoCoGAN，RGB+D-MoCoGAN の 4 つを提案し，比較検証を行った．実験より，RGBD-MoCoGAN が最も高い品質でデータを生成できることが判明した．今後の展開として，顔表情以外の動画データセットで実験を行うことや，より depth の生成に適したアーキテクチャを考案することが考えられる．

謝辞

本研究は JSPS 科研費 JP16K00231 の助成を受けたもの

ⓒ 2018 Information Processing Society of Japan



(a) RGBD-GAN (b) RGBD-VideoGAN (c) RGBD-MoCoGAN (d) RGB+D-MoCoGAN

図 5: 各モデルによる生成サンプル

です.

参考文献

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.

[4] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.

[5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[6] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016.

[7] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. *CoRR*, abs/1707.04993, 2017.

[8] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, pages 1–4. IEEE, 2010.

[9] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1031–1039. IEEE, 2017.

[10] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[12] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

ⓒ 2018 Information Processing Society of Japan