Vol.2018-CVIM-213 No.32
2018/9/21

デプスからカラーへのドメイン変換を用いたGANによる動画生成

中平有樹[1,a)]　川本一彦[2,b)]

概要：これまで動画生成のための GAN(Generative Adversarial Nets) が提案されてきたが，従来手法ではカラー動画のみを用いてモデルを訓練する．しかし，シーンのダイナミクスをより正確に捉えるためには，光学的な情報だけでなく 3 次元幾何学的な情報が重要であると考えられる．本論文では，カラー動画に加えてデプス動画を併用する GAN アーキテクチャを提案する．提案アーキテクチャの generator では，前半でデプス動画を生成し，後半でデプスからカラーへのドメイン変換を解くことでカラー動画を生成する．シーンのダイナミクスをデプスに注目してモデル化することにより，従来手法よりも自然な動きの生成を可能とした．評価実験では顔表情データセットとジェスチャデータセットでモデルを訓練し，Kernel-MMD と FID の 2 つの客観指標において，従来手法よりも高い品質で動画生成できることを示した．

キーワード：GAN，動画生成，深層学習，ドメイン変換，RGB-D データ

1. はじめに

近年，深層生成モデルを用いたデータの自動生成が大きな成功を収めている．特に GAN(Generative Adversarial Nets)[1] は静止画生成のタスクにおいて，高解像度画像の生成 [2] やクラスラベルを用いた条件付き画像生成 [3][4] に成功している．加えて，純粋な生成タスクだけでなく，画像超解像 [5]，ドメイン変換 [6][7]，テキストからの画像生成 [8] などの応用的なタスクでも軒並み高い精度を発揮している．これらの研究から今後の GAN の適用対象がより高次元で複雑なデータ，例えば動画へと移っていくことが予想される．

これまでに動画生成のための GAN アーキテクチャが複数提案されてきた [9][10][11]．これらの研究より，文字が移動する単純な動画や，人の顔表情の動画においては，高品質な動画が生成できるようになっている．しかしながら，複数の物体が含まれる一般的なシーンの動画生成においては，動いている物体の見た目が不自然になったり，物体が背景に同化するなどの問題が報告されている．動画生成は，そのような時間方向のモデル化の難しさから，画像生成に比べて成功していないと言える．より自然なシーンを生成するためには，動いている物体の領域を正確に捉え，自然な動きを付与し，同時に背景や他の物体との相互作用を学習する必要があると思われる．

我々はそのような問題を解決するために，ダイナミクスをより正確に解釈するためには，シーンの光学的な情報だけでなく，幾何学的な情報を用いるべきであると考える．これは次の 2 つの考えに基づいている．(1) 動画では動きが複数の画像（2 次元平面）で表現されているが，本質的に物体は三次元空間上で動いている．よって，平面的な光学的情報のみでシーンを正確に捉えることは難しいと言える．(2) 人間の脳内では，網膜から得られる 2 次元的な光学的情報からシーンの奥行きを推定し，考慮していることがわかっている [12]．これは外界を正しく把握するためには，三次元幾何学情報が必要不可欠であることを示唆している．本論文では，幾何学的な情報として汎用的に利用できるデプス画像を用いて動画を生成する GAN を提案する．

2. 関連研究

これまでに動画予測に関する研究が多数行われてきた．Ranzato ら [13] は動画の教師なし学習のためのベースラインとなるような，空間と時間の関係を RNN で学習してフレーム予測を行うモデルを提案した．Srivastava ら [14] は，LSTM を用いた Encoder-Decoder モデルを用いて動画のフレーム予測を行うモデルを提案した．これ

[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

らの研究を発端に現在までに数々の手法が提案されている [15][16][17][18][19][20]．しかし，動画予測のタスクは過去のフレームを入力として次フレームを予測する問題であり，0 からのビデオ生成には直接適用することができない．

2014 年に Goodfellow らによって GAN（Generative Adversarial Nets）[1] が提案された．GAN は generator と discriminator の 2 つのネットワークで構成される生成モデルである．generator はデータセットに含まれるサンプルと見分けがつかないようなサンプルを生成するように学習し，discriminator は与えられたサンプルが generator 由来かデータセット由来かを判別するように学習する．これまで，動画生成のための GAN が多数提案されてきた．Vondrick ら [9] は動画生成に初めて GAN を適用した VideoGAN を提案した．VideoGAN では，固定されたカメラで撮影した動画が動的な前景と静的な背景に分けられることを指摘した．この仮定を基に，前景にあたる動画を三次元畳み込み，背景にあたる静止画を 2 次元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るアーキテクチャを提案した．一方，Saito ら [10] は，動画認識における知見 [21][22][22] から，3DCNN は空間と時間を同等に扱うため不適切であると指摘した．そこで動画の各フレームに対応する潜在ベクトルを 1DCNN で生成する temporal generator と，画像ベースの Wasserstein-GAN[23] を用いた image generator を用意し，時間方向と空間方向を別々に生成する TGAN を提案した．Tulyakov ら [11] は，動画は内容（一貫して不変な要素）と動き（時間と共に変化する要素）に分けられると指摘し，内容と動きの 2 つの潜在ベクトルから画像フレームを生成する MoCoGAN を提案した．MoCoGAN では，動画の各フレームに対応する動きの潜在ベクトルを RNN で生成し，反対に内容の潜在ベクトルは動画内で 1 つに固定して生成することで，特別な学習なしで内容と動きを分けることに成功した．これらのアーキテクチャは，いずれも動画を時間と空間に分けてモデル化する方法に着目していると言える．その中でも，MoCoGAN の生成結果は最も品質が高く，内容を固定した上で異なる動きとするような制御も可能としている．

しかし，これらのアーキテクチャはいずれも，生成対象であるカラー動画のみを用いて訓練を行う．本論文ではより適切に動きを捉えるために，デプス動画を用いる GAN を提案する．

3. 提案手法

従来の GAN による動画生成手法では，生成の対象であるカラー動画のみを用いてモデルの訓練を行ってきた．提案手法では幾何学的な情報であるデプス動画を併用してモデルを訓練する．提案アーキテクチャは次の 2 つの考えを反映している．(1) シーンにおいて時間と共に本質的に変化しているのは幾何学的情報（位置や形，大きさ）である．よってダイナミクスのモデル化はデプスで行う．(2) シーンの幾何学的な情報が決まれば，それに対応した適切なカラーを付加することでシーンを生成できる．これはデプス空間からカラー空間へのドメイン変換問題と捉えることができる．すなわち提案モデルでは動画をデプスとカラーに分け，特にデプスでダイナミクスのモデル化を行うことで動画を生成する．このアーキテクチャを**図 1** に示す．提案モデルは主に 5 つのネットワークで構成される．それらは frame seed generator（R_M），depth image generator（G_D），color image generator（G_C），image discriminator（D_I），video discriminator（D_V）である．

はじめに，MoCoGAN のアーキテクチャに倣ってデプス動画を内容と動きに分けて生成する．frame seed generator は動画の時間変化をモデル化し，動きの潜在ベクトルを生成するネットワークである．潜在ベクトルは

$$\left[\boldsymbol{z}^{(1)},...,\boldsymbol{z}^{(T)}\right]=\left[\begin{bmatrix}\boldsymbol{z}_C\\ \boldsymbol{z}_M^{(1)}\end{bmatrix},...,\begin{bmatrix}\boldsymbol{z}_C\\ \boldsymbol{z}_M^{(T)}\end{bmatrix}\right] \tag{1}$$

のように表され，動画の各画像フレームに対応している．ここで $\boldsymbol{z_C}$ は動画の内容に対応する潜在ベクトルであり，動画内で内容（形状や大きさ）を一貫させるために共通のベクトルを用いる．潜在ベクトル $\boldsymbol{z}_C$ は，平均 0 と共分散行列 $\boldsymbol{I}_{d_C}$ の正規分布 $\mathcal{N}(0,\boldsymbol{I}_{d_C})$ から生成する．すなわち

$$\boldsymbol{z}_C\sim\mathcal{N}(0,\boldsymbol{I}_{d_C}) \tag{2}$$

と生成する．ここで，d_C は潜在ベクトル $\boldsymbol{z}_C$ の次元をあらわし，$\boldsymbol{I}_{d_C}$ は $d_C\times d_C$ の単位行列である．一方，$\boldsymbol{z}_M^{(t)},t=1,2,...,T$ は，T フレームにわたる動画の動きに対応する潜在ベクトルであり，

$$\boldsymbol{z}_M^{(t)}=R_M(\boldsymbol{z}_M^{(t-1)},\boldsymbol{\epsilon}^{(t)}),\ t=1,2,\ldots,T \tag{3}$$

のように生成される．ここで，

$$\boldsymbol{\epsilon}^{(t)}\sim\mathcal{N}(0,\boldsymbol{I}_{d_M}) \tag{4}$$

であり，d_M は潜在ベクトル $\boldsymbol{z}_M^{(t)}$ の次元をあらわす．R_M が frame seed generator であり，1 層の GRU[24] を用いる．

次に，depth image generator は潜在ベクトルからデプス動画の画像フレームを生成するネットワークである．式 (1) の潜在ベクトルを用いて，デプス動画 $\boldsymbol{x}_{\text{depth}}^{(t)}$ を

$$\boldsymbol{x}_{\text{depth}}^{(t)}=G_D(\boldsymbol{z}^{(t)}),t=1,2,...,T \tag{5}$$

と生成する．ここで，G_D が depth image generator である．G_D には 5 層の転置畳み込みの CNN を用いる．ここまでで，シーンにおいて時間と共に本質的に変化しているのは幾何学的情報であるという仮定に基づき，デプス動画の生成を通してシーンのダイナミクスをモデル化する．

次にデプス動画をカラー動画へと変換し，最終的な動画を出力する．color image generator はデプス画像をカラー

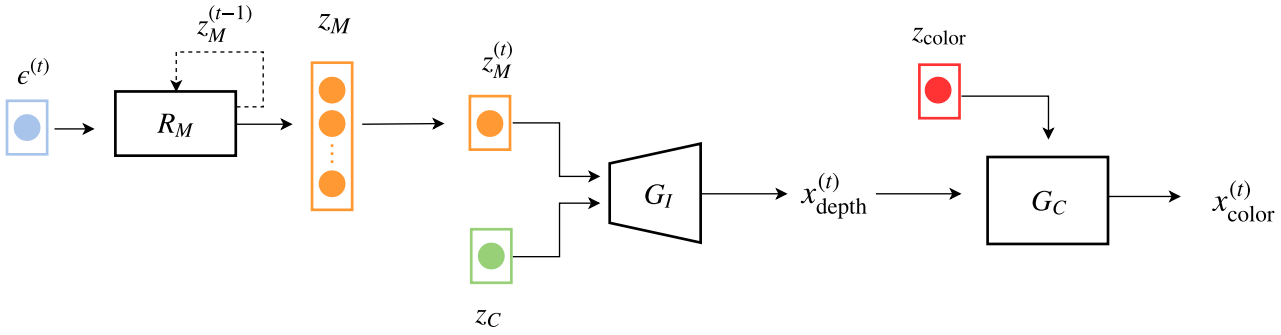


図 1: 提案モデルのアーキテクチャ

画像へと変換するネットワークである．変換は

$$\boldsymbol{x}_{\text{color}}^{(t)} = G_C\left(\boldsymbol{x}_{\text{depth}}^{(t)}, \boldsymbol{z}_{\text{color}}\right), t = 1, 2, ..., T \qquad (6)$$

のように表される．ここで，$\boldsymbol{z}_{\text{color}} \sim \mathcal{N}(\boldsymbol{z}|0, \boldsymbol{I}_{d_{\text{CR}}})$ である．G_C が color image generator である．$\boldsymbol{z}_{\text{color}}$ はデプスからカラーへの変換を確率的に変えるパラメータである．この変換はデプスからカラーへのドメイン変換の問題であると解釈できるため，pix2pix[6] のアーキテクチャに倣い，G_C には U-Net[25] を用いる．$\boldsymbol{z}_{\text{color}}$ は U-Net の encoder から得られるボトルネック特徴量に対してチャネル方向に結合してから decoder に入力するように組み込む．これにより，シーンの幾何学的な情報であるデプスに基づいてカラー動画を生成することができる．

最後に discriminator が 2 つあり，image dicriminator は動画内の画像フレームを批判することに特化し，video discriminator は動画全体を批判することに特化している．image discriminator は 5 層の 2DCNN で, 動画中のある 1 フレームを入力する．video discriminator は 5 層の 3DCNN で，動画全体を入力する．各 discriminator はカラー動画（画像）とデプス動画（画像）のペアを受け取り，それがデータセットのサンプルであるか，generator 由来のサンプルであるかを識別する．すなわち目的関数は

$$\min_{R_M, G_D, G_C} \max_{D_I, D_V} V(R_M, G_D, G_C, D_I, D_V), \qquad (7)$$

$$\begin{aligned} &V(R_M, G_D, G_C, D_I, D_V) \\ &= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}[\log D_I(\boldsymbol{x}_{\text{color}}^{(t)}, \boldsymbol{x}_{\text{depth}}^{(t)})] \\ &\quad + \mathbb{E}_{\tilde{\boldsymbol{x}} \sim p_g}[1 - \log D_I(\tilde{\boldsymbol{x}}_{\text{color}}^{(t)}, \tilde{\boldsymbol{x}}_{\text{depth}}^{(t)})] \\ &\quad + \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}[\log D_V(\boldsymbol{x}_{\text{color}}, \boldsymbol{x}_{\text{depth}})] \\ &\quad + \mathbb{E}_{\tilde{\boldsymbol{x}} \sim p_g}[1 - \log D_V(\tilde{\boldsymbol{x}}_{\text{color}}, \tilde{\boldsymbol{x}}_{\text{depth}})]. \end{aligned}$$

である．ここで $\boldsymbol{x}_{\text{color}}$，$\boldsymbol{x}_{\text{depth}}$ はデータセットの動画サンプルを，$\tilde{\boldsymbol{x}}_{\text{color}}$，$\tilde{\boldsymbol{x}}_{\text{depth}}$ は生成された動画サンプルを示す．D_I は image discriminator, D_V は video discriminator, p_g は generator による生成サンプルの分布，p_{data} はトレーニングデータの分布を示す．

4. 評価実験

本節では，人の顔表情データセットとジェスチャデータセットを用い，提案手法と MoCoGAN[11] との比較を行う．

4.1 データセット

実験に用いたデータセットは次の 2 つである．

- **顔表情データセット**: MUG Facial Expression Database[26] を用いる．これは男女 86 人が様々な表情を演じるデータセットである．表情には怒り，嫌悪感，恐怖，喜び，悲しみ，驚きの 6 種類がある．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いる．本データセットはデプスデータが収録されていないため，顔の三次元復元手法 [27] を用いてデプスデータを予測し生成した．また画像サイズを 64 × 64 に縮小し，各動画から長さ 16 のサブシーケンスを抜き出して用いる．
- **ジェスチャデータセット**: ChaLearn LAP IsoGD Dataset[28] を用いる．これは 21 人の男女が様々なハンドジェスチャを行う RGB-D 動画データセットである．ジェスチャには 249 のラベルが存在し，合計 47933 サンプルが収録されている．背景や手の動きによるオクルージョンが発生するため，顔表情データセットよりも複雑なシーンの動画データであるといえる．画像サイズを 64 × 64 に縮小し，各動画から長さ 16 のサブシーケンスを抜き出して用いた．

4.2 実験条件

最適化手法には ADAM[29]（学習率 0.0002．モーメンタム 0.5 と 0.999）を用い，学習を安定させるために D_I，D_V の各層の入力に $\mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$ のノイズを付加した [30]．潜在ベクトルの次元はそれぞれ $d_C = 40$，$d_M = 10$，$d_{\text{CR}} = 50$ とした．

4.3 主観評価

各データセットの生成結果を図 2，3 に示す．提案手法

ⓒ 2018 Information Processing Society of Japan

表 1: Kernel MMD による評価

	MoCoGAN	proposed model
顔表情データセット	368 ± 0.518	$\mathbf{340 \pm 0.359}$
ジェスチャデータセット	275 ± 1.04	$\mathbf{235 \pm 0.552}$

表 2: FID（10^{-3}）による評価

	MoCoGAN	proposed model
顔表情データセット	280 ± 2.02	$\mathbf{245 \pm 1.46}$
ジェスチャデータセット	580 ± 5.61	$\mathbf{534 \pm 4.02}$

は，MoCoGAN と同程度に高い画質の動画が生成できることがわかる，特にジェスチャデータにおいて，MoCoGAN よりも人の外見や手の動きを自然に生成できることがわかった．例えば，MoCoGAN の生成サンプルの中には，人の手が 3 本以上存在したり（**図 4a**），人が 2 人以上映ってしまう（**図 4b**）といった現象が起こったが，提案手法では見られなかった．これは，シーンのデプスを考慮することによって，背景と人，動きのある手の部分をよりうまく区別できたためと考えられる．

4.4 客観評価

提案モデルと MoCoGAN を客観評価指標で比較した．GAN を評価する上で広く用いられる指標に Inception Score[30] があるが，最近の研究から適切な指標ではない可能性が示唆されている [31][32]．よって本実験では，[31] で適切な評価指標であると結論付けられている Kernel MMD[33] と Inception Score の次によく用いられる FID(Fréchet Inception Distance)[34] を採用する．

4.4.1 Kernel MMD

Kernel MMD(Maximum Mean Discrepancy)[33] は 2 つの確率分布の距離を計算する指標であり，

$$\mathrm{MMD}(\mathbb{P}_r, \mathbb{P}_g) = \left(\mathbb{E}_{\substack{x_r, x_r' \sim \mathbb{P}_r, \\ x_g, x_g' \sim \mathbb{P}_g}} \left[k(\boldsymbol{x}_r, \boldsymbol{x}_r') - 2k(\boldsymbol{x}_r, \boldsymbol{x}_g) + k(\boldsymbol{x}_g, \boldsymbol{x}_g') \right] \right)^{\frac{1}{2}} \quad (8)$$

のように表される．ここで，$\mathbb{P}_r$，$\mathbb{P}_g$ はそれぞれ学習データと generator の分布を指し，k はベクトルの内積とする．Kernel MMD の値は 2 つの確率分布の距離（非類似度）を示すため，小さければ小さいほど良い．指標を算出する際は，それぞれの確率分布から 100000 回ランダムにサンプリングを行って MMD を計算し，その平均値をスコアとする．また，$\boldsymbol{x}_r, \boldsymbol{x}_r', \boldsymbol{x}_g, \boldsymbol{x}_g'$ は動画のピクセルデータをそのまま用いるのではなく，UCF-101[35] によって訓練された C3D[36] に入力して得られる中間層の特徴量を用いる．これは深層ニューラルネットワークから得られる中間層の特徴を用いたほうが，より正確に分布間類似度を計測できるためである [31]．

評価結果を**表 1** に示す．両データセットで提案手法が MoCoGAN を上回る結果となった．

4.4.2 FID

FID(Fréchet Inception Distance)[34] は GAN の生成画像の品質を評価するために提案された指標の 1 つであり，先の指標と同様に 2 つの確率分布間の距離を定義するものである．FID はそれぞれの確率分布を多変量ガウス分布として捉え，その平均と分散を用いて

$$\mathrm{FID}(\mathbb{P}_r, \mathbb{P}_g) = ||\boldsymbol{\mu_r} - \boldsymbol{\mu_g}||_2^2 + \mathrm{Tr}(\boldsymbol{\Sigma_r} + \boldsymbol{\Sigma_g} - 2(\boldsymbol{\Sigma_r}\boldsymbol{\Sigma_g})^{\frac{1}{2}}) \quad (9)$$

のように表される．ここで $\mathbb{P}_r$，$\mathbb{P}_g$ はそれぞれ学習データと generator の分布，$\boldsymbol{\mu_r}$，$\boldsymbol{\Sigma_r}$ は $\mathbb{P}_r$ の平均ベクトルと共分散行列，$\boldsymbol{\mu_g}$，$\boldsymbol{\Sigma_g}$ は $\mathbb{P}_g$ の平均ベクトルと共分散行列を示す．FID の値は 2 つの確率分布の距離（非類似度）を示すため，小さければ小さいほど良い．指標を算出する際は，それぞれの確率分布から得られる 3000 サンプルを用いて計算する．また，動画のピクセルデータをそのまま用いるのではなく，Kernel-MMD と同様に C3D に入力して得られる中間層の特徴量を用いる．

評価結果を**表 2** に示す．両データセットで提案手法が MoCoGAN を上回る結果となった．

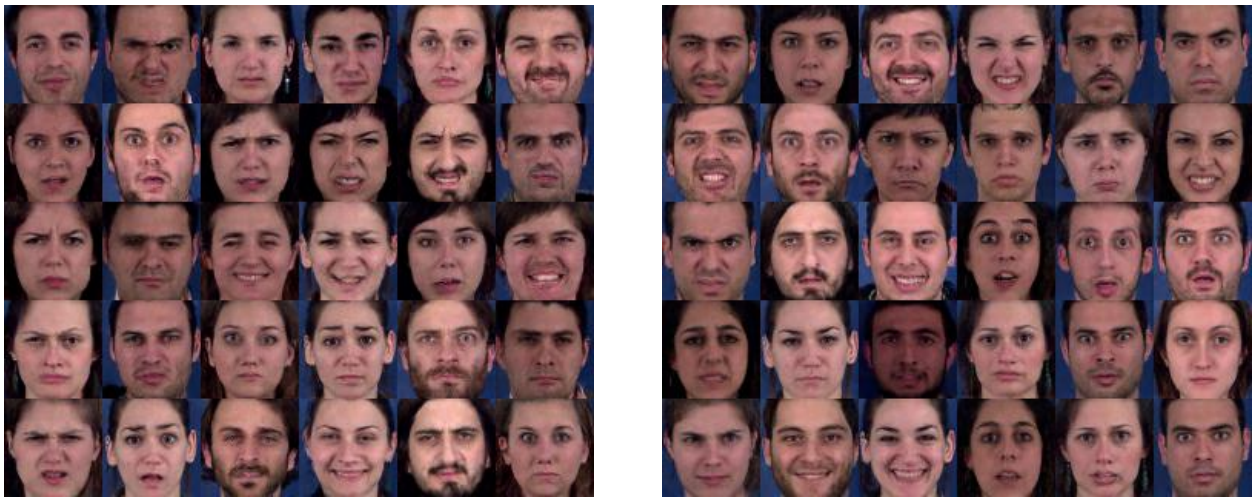
4.5 考察

主観評価では動画の個々のフレームの画質に有意な違いは見られなかったが，客観評価においては提案手法が MoCoGAN を大きく上回る結果となった．デプスを考慮することによって動画のダイナミクスやシーン内容の自然性が向上したと考えられる．また，顔表情データセットでは顔の三次元復元手法 [27] によって推定されたデプスを用いたにもかかわらず，精度が向上した．これは，データの計測に高いコストがかかる場合にも，予測ベースのデータを用意することで精度を改善できることを示唆している．

一方で，MoCoGAN などでは顔表情データを生成した際に，データセットに含まれる 2 人の顔の中間を取ったようなサンプルや，顔の右半分と左半分で見た目や動きが異なるようなサンプルがいくつも見られた．このような生成結果は人の顔として不自然である一方，生成モデルがよく汎化していることを示していると言える．しかし，提案手法の生成結果にはそのようなサンプルがほとんど見られなかった．このことから，提案手法はより自然なサンプルを生成できる一方で，データの多様性が低い可能性があると考えられる．

最後に，提案手法では生成されたデプス動画に対して任意の色を付加できるよう，color image generator に $\boldsymbol{z}_{\mathrm{color}}$ を与えていた．しかし，同じデプス動画に異なる $\boldsymbol{z}_{\mathrm{color}}$ を与えてもネットワーク内では単に無視され，出力されるカラー動画が同一になった．これは [6] でも言及されている現象であり，目的関数の異なる本モデルでも同様に起きた

ⓒ 2018 Information Processing Society of Japan



(a) MoCoGAN　(b) 提案手法

図 2: 顔表情データセットの生成結果



(a) MoCoGAN　(b) 提案手法

図 3: ジェスチャデータセットの生成結果

ことから，U-Net の skip-connection 構造に起因するものであると考えられる．今後，color image generator のモデルを改良することで，例えば顔のデータであれば，A さんの骨格に B さんの顔を付与したようなサンプルを生成できるようになり生成データの多様性が高まると考えられる．

5. おわりに

動画生成を行う上で重要なのはシーンのダイナミクスの理解である．我々は動画に写る物体やシーンが本来は三次元空間にあることを踏まえ，三次元の幾何学的情報であるデプスを用いて動画を生成する手法を提案した．実験より，GAN による動画生成の state-of-the-art である MoCoGAN を超える精度で動画を生成できることを示した．本研究の結果から，シーンの理解には光学的情報だけでなく幾何学的情報も考慮すべきであること，幾何学的情報としてデプスは有用であることが示唆される．今後も，カラーとデプスを併用してより適切にシーンをモデル化できる手法を検討していく必要がある．

謝辞

本研究は JSPS 科研費 JP16K00231 の助成を受けたものです．

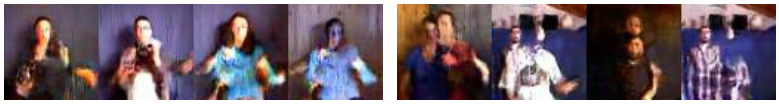
参考文献

[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] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations*, 2018.

[3] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*,

ⓒ 2018 Information Processing Society of Japan



(a) 人の手が 3 本以上映る例　(b) 人が 2 人以上映る例

図 4: 従来手法で見られた生成結果

2014.

[4] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[5] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 105–114, 2017.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. July 2017.

[7] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017.

[8] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *CoRR*, abs/1605.05396, 2016.

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

[10] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[11] Xiaodong Yang Sergey Tulyakov, Ming-Yu Liu and Jan Kautz. Mocogan: Decomposing motion and content for video generation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Ken-Ichiro Tsutsui, Hideo Sakata, Tomoka Naganuma, and Masato Taira. Neural correlates for perception of 3d surface orientation from texture gradient. *Science*, 298(5592):409–412, 2002.

[13] Marc'Aurelio Ranzato, Arthur Szlam, Joan Bruna, Michaël Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *CoRR*, abs/1412.6604, 2014.

[14] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 843–852, 2015.

[15] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2863–2871. Curran Associates, Inc., 2015.

[16] Yann LeCun Michael Mathieu, Camille Couprie. Deep multi-scale video prediction beyond mean square error. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[17] Nal Kalchbrenner, Aäron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[18] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 64–72. Curran Associates, Inc., 2016.

[19] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 91–99. Curran Associates, Inc., 2016.

[20] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1878–1889. Curran Associates, Inc., 2017.

[21] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc., 2014.

[22] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[23] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[24] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS Deep Learning Workshop*, 2014.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

ⓒ 2018 Information Processing Society of Japan

[26] 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.

[27] 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.

[28] Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera, and Stan Z Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64, 2016.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[30] 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.

[31] Gao Huang, Yang Yuan, Qiantong Xu, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks, 2018.

[32] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[33] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

[34] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017.

[35] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[36] 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