

# ダイナミクススタイル変換を用いた自然風景画像の動画化

栗崎 一真<sup>1</sup> 川本 一彦<sup>2</sup>

**概要:** 本研究では、動画コンテンツ作成支援に向けた自然風景画像の動画化法を提案する。提案手法では、自然風景に含まれる草原や滝などの領域の動的変化をダイナミクススタイル変換により与える。ダイナミクススタイル変換は、テクスチャの動的変化だけでなくスタイル変換も学習により与えることができる。一方で、スタイル損失では位置情報が失われてしまう。そこで、スタイル損失と二乗損失を組み合わせた損失関数を導入する。3つの自然風景画像の動画化の自然さに関して、被験者21名による主観評価を実施したところ、川や滝のような流体と麦畑や雲のような領域では、スタイル損失と二乗損失のバランスを変化させたほうが良いことが示された。さらにスタイル変換しつつ動画化するモデルを提案し、動画生成を多様なものに発展させた。

## 1. はじめに

近年、深層学習を用いた画像・動画コンテンツの生成の研究が関心を集めている。その例として敵対性ネットワークGAN[1]による動画生成やスタイル変換による画風変換がある。本研究では自然風景画像の背景領域（水面や空など）に動的変化を与える動画化法を提案する。自然風景画像には様々な領域が含まれる。これらの自然風景画像の領域は空間的および時間的に繰り返しのパターンがあり、2次元信号のテクスチャパターンのダイナミクスで記述できる[2]。

提案手法では、領域ごとにダイナミクステクスチャを付与し動画化する。ダイナミクステクスチャの生成には古典的にdynamics texture法[3]があるが、最近ではstyle transferを拡張したdynamics style transfer (DST)[4]が提案されている。提案手法ではDSTをベースにしつつ、見えの不自然さを減らすために新たな損失を導入し、DSTよりも自然な動的背景を生成できるようにする。DSTは見えと動きをtwo-streamで学習するが、そこにスタイルを学習するstreamを追加し、three-streamで見え、動き、スタイルを同時に学習するモデルも提案する。このモデルにより見え、スタイル、動きについて3つ同時に学習することが可能となり、多様な動画の生成を実現する。

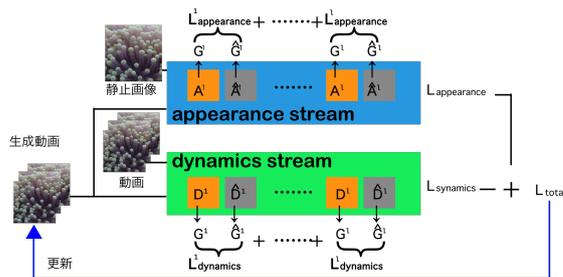


図1 dynamics style transfer アーキテクチャ

## 2. ダイナミクススタイル変換

Dynamics style transfer はダイナミクステクスチャ生成手法の1つである[4]。このモデルは図1に示すようにappearance stream と dynamics stream からなる。

Appearance stream からは見えの特徴を学習し、dynamics stream からは動きの特徴を学習する。普通の学習ではネットワークの重みを更新することで最適化するが、この手法ではネットワークの重みは更新しない。更新するのは入力画像である。見えのターゲットとなるテクスチャ画像を appearance stream に、動きのターゲットとなるテクスチャ動画を dynamics stream に入力し、見えと動きの特徴をモデルが獲得する。

### 2.1 Appearance stream

プレトレーニングされたモデルに入力動画とターゲットテクスチャ動画を入力し、それぞれ見えの特徴を抽出し損

<sup>1</sup> 千葉大学 大学院融合理工学府  
Graduate School of Science and Engineering,  
Chiba University  
<sup>2</sup> 千葉大学大学院工学研究院  
Graduate School of Engineering, Chiba University  
1-33, Yayoicho, Inage-ku Chiba-shi, Chiba, 263-8522 Japan

失を計算する。モデルは VGG19[5] である。VGG19 は 19 層からなる CNN で物体認識などに用いられる。Appearance stream ではターゲットテキストチャ動画と入力動画を VGG19 に入力し、中間層で特徴マップを作成する。ここで得られた特徴マップでターゲットテキストチャ動画と入力動画で損失をとる。損失は Gatys らが提案したスタイル損失  $L_{style}$  [6] を用いる。スタイル損失式を  $L_{style}$  を式 (1) に示す。

$$L_{style} = \frac{1}{L_{app}T_{out}} \sum_{t=1}^{T_{out}} \sum_l \left\| G^l - \hat{G}^{lt} \right\|_F^2 \quad (1)$$

ここで、 $\|\cdot\|_F$  はフロベニウスノルム、 $L_{app}$  はレイヤ数、 $T_{out}$  は生成フレーム数、 $l$  はレイヤ番号、 $t$  はフレーム番号、 $M$  は画素数、 $G^{lt}$  はレイヤ番号  $l$  とフレーム番号  $t$  のグラム行列である。スタイル損失  $L_{style}$  はチャンネル間の相関を考慮する損失である。ターゲットテキストチャ動画と入力動画のレイヤごとの特徴マップからそれぞれグラム行列  $\hat{G}$  と  $G$  を計算し、差を損失とする。グラム行列の計算には VGG の conv1\_1, pool1, pool2, pool3 そして pool4 レイヤの特徴マップを用いる。グラム行列の  $i$  行  $j$  列要素は式 (2) および (3) で表される。

$$G_{ij}^l = \frac{1}{TN_l M_l} \sum_{t=1}^T \sum_{k=1}^{M_l} A_{ik}^{lt} A_{jk}^{lt} \quad (2)$$

$$\hat{G}_{ij}^{lt} = \frac{1}{N_l M_l} \sum_{k=1}^{M_l} \hat{A}_{ik}^{lt} \hat{A}_{jk}^{lt} \quad (3)$$

ここで、 $l$  はレイヤ番号、 $t$  はフレーム番号、 $T$  は入力フレーム数、 $N$  はチャンネル数、 $M$  は画素数、 $A_{ab}^{lt}$  はレイヤ番号  $l$  とフレーム番号  $t$  の特徴マップ  $A$  のチャンネル番号  $a$  画素インデックス  $b$  の画素値である。

以上のようにして得られたスタイル損失  $L_{style}$  を見え損失  $L_{appearance}$  とする。

$$L_{appearance} = L_{style} \quad (4)$$

## 2.2 Dynamics stream

Dynamics stream においても appearance stream と同じようにプレトレーニングされたモデルに入力動画とターゲットテキストチャ動画を入力する。Dynamics stream のアーキテクチャを図 2 に示す。Dynamics stream は空間情報と時間情報の特徴を抽出することに特化した spacetime-oriented energy model[7][8] を使用する。このモデルは UCF101 dataset[9] でプレトレーニングされている。プレトレーニングでは生成された optical flow と flow estimator[10] で生成された optical flow の  $L_2$  損失を用いて学習する。損失は式 (5) で示され appearance stream の  $L_{appearance}$  と基本的に同じである。

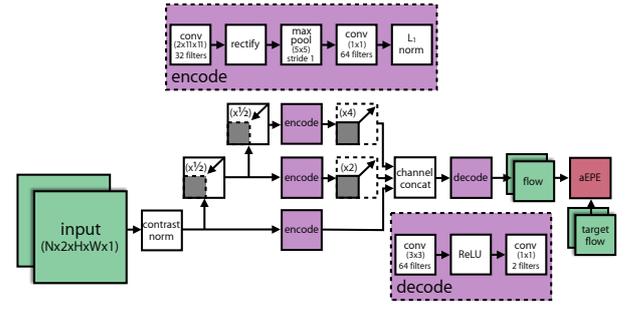


図 2 dynamics stream のアーキテクチャ [4]

$$L_{dynamics} = \frac{1}{L_{app}T_{out} - 1} \sum_{t=1}^{T_{out}-1} \sum_l \left\| G^l - \hat{G}^{lt} \right\|_F^2 \quad (5)$$

ここで、 $L_{app}$  はレイヤ数、 $T_{out}$  は生成フレーム数、 $l$  はレイヤ番号、 $t$  はフレーム番号、 $M$  は画素数、 $G^{lt}$  はレイヤ番号  $l$  とフレーム番号  $t$  のグラム行列である。Appearance stream との違いは optical flow を計算するためフレーム数が 1 つ少なくなっているところである。グラム行列の計算には dynamics stream のピラミッド構造の後の concat レイヤの特徴マップを用いる。

## 2.3 動画生成

Appearance stream と dynamics stream からそれぞれ  $L_{appearance}$  および  $L_{dynamics}$  を計算しその和を全体の損失とする。

$$L_{total} = \lambda_1 L_{appearance} + \lambda_2 L_{dynamics} \quad (6)$$

ここで  $\lambda_1$  および  $\lambda_2$  は見え損失と動き損失のバランスをとるための係数である。この損失を最小化する方向に入力動画を更新していくことで動画を生成する。

## 3. 提案手法

### 3.1 自然風景画像の動画化

自然風景画像には異なるダイナミクスをもつテキストチャが含まれている。本研究ではセマンティックセグメンテーションを使い自然風景画像を同じダイナミクステキストチャの領域に分けてそれぞれを DST を用いて動画化する。生成の工程を図 3 に示す。

見えの不自然さを軽減するために DST の損失関数を次のように変更する。DST のスタイル損失  $L_{style}$  は特徴マップのグラム行列を用いて計算されるため、画像の空間的位置情報が失われてしまう。自然風景画像において位置情報が失われることは見えに不自然さを与える。位置情報は、式 (7) の  $L_2$  損失を加えることで失われることを抑制することができる。

$$L_2 = \frac{1}{TN_l} \sum_{t=1}^T \sum_{i,j} (A_{ij}^{lt} - \hat{A}_{ij}^{lt})^2 \quad (7)$$

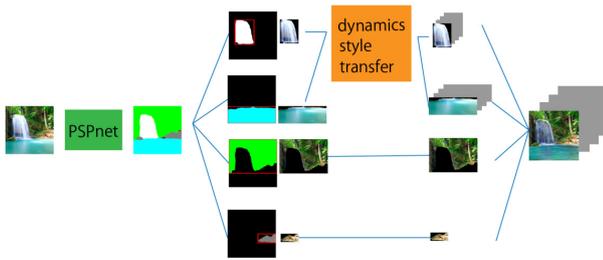


図 3 提案モデル

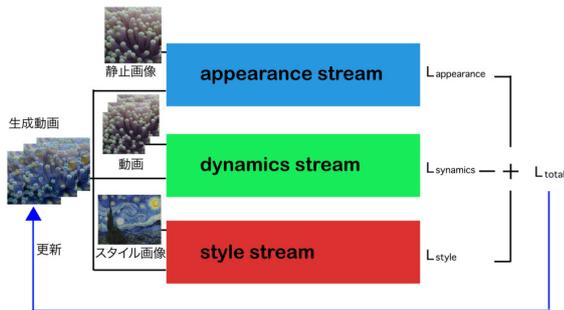


図 4 three-stream 生成モデル

ここで、 $l$  はレイヤ番号、 $t$  はフレーム番号、 $T$  はフレーム数、 $N$  はチャンネル数、 $M$  は画素数、 $A_{ab}$  は特徴マップ  $A$  のチャンネル番号  $a$ 、画素インデックス  $b$  のピクセル値である。

提案手法では、式 (8) に示すように、 $L_{style}$  損失だけでなく  $L_2$  損失を加えた見え損失  $L_{appearance}$  を導入する。

$$L_{appearance} = \lambda_3 L_{style} + \lambda_4 L_2 \quad (8)$$

ここで  $\lambda_3$  および  $\lambda_4$  はスタイル損失と  $L_2$  損失のバランスをとるための係数である。

### 3.2 Three stream ダイナミクススタイル変換

DST の生成の多様性を上げるため style stream というスタイルを学習する stream を加える。Three-stream で学習することでスタイル変換をしつつ動画化する手法を提案する。Style stream のネットワークには appearance stream と同じ VGG19 を使用する。このネットワークの概略を図 4 に示す。動画生成過程は two-stream のときと同じで、入力動画を更新することで動画を生成する。Three-stream ダイナミクススタイル変換の損失は、次のように 3 つの stream に関する損失の和として定義する

$$L_{total} = \lambda_5 L_{appearance} + \lambda_6 L_{dynamics} + \lambda_7 L_{style} \quad (9)$$

ここで  $\lambda_5$ 、 $\lambda_6$  および  $\lambda_7$  は見え損失、動き損失およびスタイル損失のバランスをとるための係数である。



図 5 動画化する静止画像

## 4. 実験

### 4.1 セマンティックセグメンテーション

セマンティックセグメンテーションには Pyramid Scene Parsing Network (PSPNet) [11] を用いた。トレーニングには ADE20K [12] を使用した。ADE20K は 150 以上のクラスを持つが本研究では自然風景のみを扱うため、推測ではすべてのクラスは必要ない。そこで水、空、滝、草木、地面、岩のみを推測し、それ以外はその他に分類して、計 7 クラスに推測するようにした。

### 4.2 生成結果

まず、two-stream の DST で生成した自然風景動画について示す。動画化する静止画像として図 5 の 3 つの画像を用意する。それぞれ水、空、草、および滝の領域を動画化の対象とする。動きのターゲットとなる動画は水面、雲、揺れる草、および滝の動画を用意し、図 6 に示す。DST の appearance stream に図 5 そして dynamics stream に図 6 を入力し動画を生成する。

### 4.3 評価実験

生成された動画を主観評価実験で評価する。この実験では式 (8) の  $L_{appearance}$  の  $L_2$  と  $L_{style}$  が生成結果の自然さに与える影響を調べる。自然風景画像を 3 種類用意する。それぞれの画像に対して 2 種類の提案した見え損失式 (8) で動画を生成する。1 つ目の損失は  $\lambda_3 = 1$ 、 $\lambda_4 = 0$  とした損失で、すなわち  $L_2$  損失である。2 つ目の損失は  $\lambda_3 = 1$ 、 $\lambda_4 = 0.1$  とした損失である。この係数の値は試行錯誤で決定した。 $L_{style}$  を 0.1 以上大きくすると明らかに見えが崩れて不自然になる。

生成された動画に対して、20 代の 21 名の被験者に自然さを 5 段階で評価してもらおう。1 が自然さの最低評価、5 が自然さの最高評価とする。それぞれの平均点をまとめたものを表 1 に示す。

### 4.4 考察

表 1 から麦畑と雲の画像以外で見え損失は  $\lambda_3 = 1$ 、 $\lambda_4 = 0$  とした  $L_2$  のみの損失の方が評価が高い傾向にある。しかし、麦畑と空の画像に関しては  $\lambda_3 = 1$ 、 $\lambda_4 = 0.1$  とした損失の方が高い結果となった。これは麦や雲といった細部の形があるものは  $L_2$  損失で位置が固定され、動きがなくなるため不自然に見えてしまうからであると考えられる。このように見え損失の  $L_2$  と  $L_{style}$  の重みによる影響は画像に

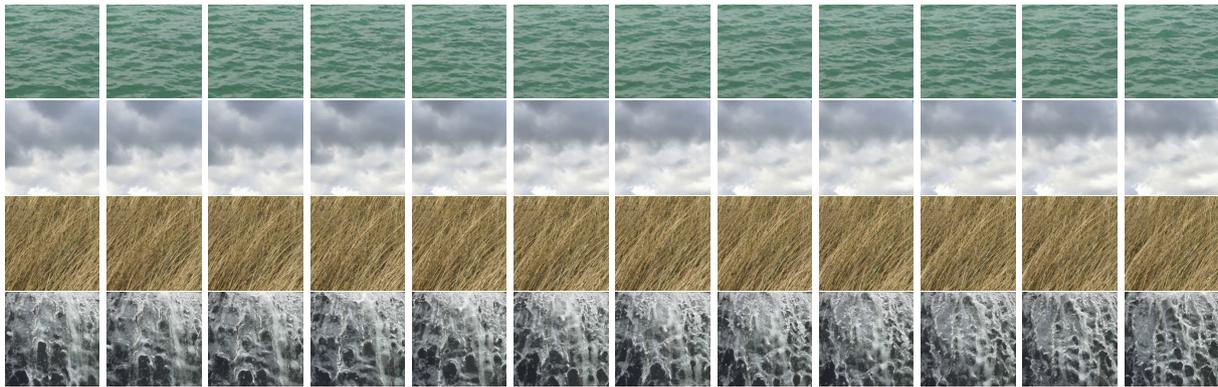


図 6 動きのターゲットとなる動画

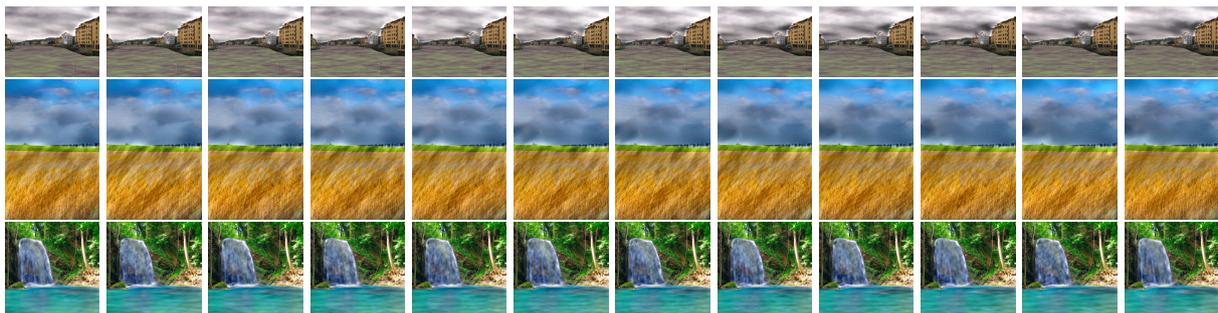


図 7 Two-stream モデルから生成された動画



図 8 スタイル画像



図 9 Three-stream モデルで動画化する静止画像

表 1 主観評価実験

Appearance	川と雲			麦畑と雲			池と滝		
	見え	動き	全体	見え	動き	全体	見え	動き	全体
$\lambda_3 = 1, \lambda_4 = 0$	2.5	3.0	2.7	3.0	3.4	3.1	3.7	3.8	3.8
$\lambda_3 = 1, \lambda_4 = 0.1$	2.0	2.2	2.0	3.5	3.2	3.3	2.0	2.6	2.4

よって異なる。  $L_2$  損失が見え損失には必要であるが、その最適な重みは画像によって異なる。適切な重みを見つけることが自然な動画生成には必要である。

#### 4.5 Three-stream で生成した動画

動画化する静止画像として図 9 の 3 つの画像を用意する。それぞれ空の部分の動画化の対象とする。動きのターゲットとなる動画は図 6 の雲の動画を用意する。スタイルのターゲットとして図 8 (ゴッホの絵画) を用意する。式 (9) の損失の係数は  $\lambda_5 = 1, \lambda_6 = 10, \lambda_7 = 0.1$  とした。これは試行錯誤で決定した。

図 6 の雲の動画と図 8 のスタイル画像から動きとスタイルを学習し、図 9 の静止画像を動画化したものが図 10 である。

## 5. おわりに

本研究では DST を用いて自然風景を動画化する方法を提案した。自然風景をセマンティックセグメンテーションしてシーンごとに分け、それぞれに合った動画を DST に入れ、再合成する手法を提案した。さらに dynamics style transfer の損失に  $L_2$  損失を加えることによって見えの位置情報を明示的に学習させ、生成動画の質を向上させた。生成の多様性を上げるために three-stream で学習してスタイル変換しつつ動画化する手法も提案した。見え、動き、スタイルを 3 つ同時に学習することで生成の多様性を向上させた。

謝辞 本研究は JSPS 科研費 JP16K00231, JP19K12039 の助成を受けたものです。

#### 参考文献

- [1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative Adversarial Nets, *Advances in Neural Information Processing Systems 27*, pp. 2672–2680 (2014).
- [2] Doretto, G., Chiuso, A., Wu, Y. N. and Soatto, S.: Dynamic textures, *International Journal of Computer Vision*, Vol. 51, No. 2, pp. 91–109 (2003).
- [3] Soatto, S., Doretto, G. and Wu, Y. N.: Dynamic textures, *IEEE International Conference on Computer Vision. (ICCV)*, Vol. 2, pp. 439–446 (2001).



図 10 Three-stream から生成された動画

- [4] Tesfaldet, M., Brubaker, M. A. and Derpanis, K. G.: Two-Stream Convolutional Networks for Dynamic Texture Synthesis, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6703–6712 (2018).
- [5] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, Vol. abs/1409.1556 (2014).
- [6] Gatys, L., Ecker, A. S. and Bethge, M.: Texture Synthesis Using Convolutional Neural Networks, *Advances in Neural Information Processing Systems 28*, pp. 262–270 (2015).
- [7] Derpanis, K. G. P. and Wildes, R.: Spacetime Texture Representation and Recognition Based on a Spatiotemporal Orientation Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 6, pp. 1193–1205 (2012).
- [8] Simoncelli, E. P. and Heeger, D. J.: A model of neuronal responses in visual area MT, *Vision Research*, Vol. 38, No. 5, pp. 743–761 (1998).
- [9] Soomro, K., Zamir, A. R. and Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, *CoRR*, Vol. abs/1212.0402 (2012).
- [10] Revaud, J., Weinzaepfel, P., Harchaoui, Z. and Schmid, C.: EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow, *CoRR*, Vol. abs/1501.02565 (2015).
- [11] Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J.: Pyramid Scene Parsing Network, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890 (2017).
- [12] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A. and Torralba, A.: Semantic Understanding of Scenes through the ADE20K Dataset, *CoRR*, Vol. abs/1608.05442 (2016).