RamenStyleAsYouLike: 領域毎のスタイル特徴の融合による画像生成

趙 宰亨[1,a)]　岡本 開夢[1,b)]　下田 和[1,c)]　柳井 啓司[1,d)]

概要

近年、深層学習技術の発展により画像の生成・変換の技術が盛んに研究されている。Pix2Pix や SPADE(spatially-adaptive normalization) を用いることで、手書きの領域マスク画像から実画像を生成することが容易に可能となっている. そこで昨年の MIRU では、我々が独自に構築したピクセル単位でラーメンのトッピングのアノテーションが付与された UEC-Ramen555 [*1]を利用して、手書きのラーメンマスク画像からラーメン画像を生成する RamenAsYouLike のデモ発表を行った. しかしながら、丼のスタイルや背景の色をユーザが指定することができず、セグメンテーションモデルと組み合わせた場合、生成される画像は元の画像とは雰囲気の異なるラーメンとなってしまうという問題点があった.

そこで、本研究では、実画像の各領域からスタイル特徴を抽出し、生成時にはマスク画像に加えて各マスク領域のスタイル特徴を合わせて与えることで、自由に領域のスタイルを制御可能とする手法を提案する。これによって、スープ、丼、チャーシューなどのスタイルをユーザが自由に指定することが可能となり、多数のラーメン画像からユーザが好みのトッピングスタイルを選んで、その特徴量を元に「究極のラーメン画像」を生成することが可能となった。

提案手法を Web ベースのシステムとして実装した「RamenStyleAsYouLike」[*2]のオンラインデモを MIRU で期間中限定で公開するので、ぜひ「究極の一杯」の生成を体験して頂きたい.

1. はじめに

近年、Web 上のブログや Twitter と Instagram のようなソーシャルネットワークサービス（SNS）などに大量の画像がアップロードされている。ユーザーは Web 上に画像をアップロードする時、より魅力的な画像をアップロードしたいから、または様々な目的で魅力的な画像を作りたいと考える。そのように画像を魅力的に作るために、元画像を編集したり、異なる画像を用いて画像を合成する場合がある。しかし、そのような画像を編集または合成する作業は熟練した画像編集のスキルと多くの時間を必要とする難しい作業である。

一方近年、深層学習の発展により様々な研究やタスクの精度が飛躍的に向上された。特に深層学習を用いた GAN(Generative Adversarial Networks) という高品質の画像生成や変換する強力なフレームワークの登場により、様々な画像生成や変換に関する研究が盛んに行い、画像生成技術が大幅に改善され、より高品質の画像を生成することができるようになった。しかし、一般的な GAN のネットワークは生成画像の属性の制御ができなかったので、GAN の構造に条件を付与して生成画像を制御する Conditional GAN(cGAN)[7] を提案することで、生成画像を制御することができるようになった。Image-to-image 変換はソースドメインからターゲットドメインへのマッピングを学習して画像から画像変換を行う手法として、生成画像の属性が制御できる cGAN のネットワークに入力画像を潜在ベクトルにエンコーディングする Encoder を追加して、画像を変換を行う手法である。

GAN を基にした様々な研究には、人の顔画像、数字画像、風景画像または都市景観画像のデータセットが幅広く使用されている。一方、GAN を使用して食品の画像を生成または変換するタスクはかなり少ない。また、「ラーメン」は日本で最も人気のある食べ物であり、アジア、アメリカ、ヨーロッパなど世界中で人気のある日本料理である。

本研究では、ピクセル単位のラベルが付けられたラーメン画像が含まれているラーメン画像データセットと Image-to-image ネットワークを用いて、ユーザーがスケッチしたマスク画像を基にして各要素の形状を制御したリアルなラーメン画像生成を行う。しかし、生成結果がユーザーが希望するスタイルと違う画像が生成され、生成画像のスタイルが制御できない問題がある。そのため、各要素のスタイル特徴を抽出する Style encoder を追加することで、スタイルを反映したラーメン画像生成を行う。

2. 関連研究

GAN は Generator と Discriminator の 2 つのネットワークから構成され、Discriminator は実画像と合成画像を判別、Generator は Discriminator が実画像と合成された画像を判別できないようによりリアルな画像を生成することを目的にする画像生成のフレームワークである。GAN のネットワークは様々な応用が可能だったので、生成画像の品質を向上する研究及び多様なタスクに GAN を基にする応用研究が盛んに行っている。GAN は正規分布などの潜在変数をサンプリングして画像を生成するが、生成画像を

[1] 電気通信大学 情報学専攻
[a)] cho@mm.inf.uec.ac.jp
[b)] okamoto-ka@mm.inf.uec.ac.jp
[c)] shimoda-k@mm.inf.uec.ac.jp
[d)] yanai@cs.uec.ac.jp

[*1] https://mm.cs.uec.ac.jp/UEC-Ramen555/
[*2] https://mm.cs.uec.ac.jp/RamenStyleAsYouLike/

制御ができない。Conditional GAN(cGAN)[7] は GAN の構造に条件を付与して属性を制御する画像生成ができるようになった。

Image-to-image 変換はソースドメインからターゲットドメインへのマッピングを学習することが目的として、cGAN の構造に Encoder を追加することで、画像変換ができるようになった。Pix2pix[4] は Encoder-Decoder 構造を U-Net[9] というネットワーク構造を使用し、互いに対応するペア画像を使用して画像変換を実現した。UNIT[6] と CycleGAN[11] では、ペア画像を学習し使用する Pix2pix とは異なり、互いに対応されない画像を用いて、画像変換を提案した。しかし、複数のドメイン間のマッピングを学習するためには、すべてのドメインに対してモデルが必要になるの問題がある。例えば、k 個のドメインを学習するためには、$k(k-1)$ 個のモデルの学習が必要がある。StarGAN[2] では、1 つのモデルで複数のドメイン間の Image-to-image 変換を提案した。MUNIT[3] と DRIT[5] は画像の各要素のコンテンツとスタイルなどの特徴を分離させ、その特徴を特徴別に制御して、画像変換を行う Disentangled representation による画像変換を提案した。Pix2pixHD[10] は、最大 256 × 256 の画像サイズまでが合成ができない制約がある Pix2pix を改善して、2048 × 1024 サイズの高解像度の画像の合成が可能になる。SPADE[8] では、正規化した後に損失される情報を防ぎ、より良い品質の画像生成のために、SPADE(spatially-adaptive normalization) という正規化レイヤーを提案する研究である。また、SEAN[12] は SPADE を拡張し領域マスク毎にスタイルを指定して画像生成を行う手法を提案しており、本研究と目的がほぼ同じである同時研究である。

本研究では、実画像とマスク画像のペアからなる画像データセットから高品質の Mask-to-Image 変換が可能である SPADE を拡張することによって、スタイルを考慮した画像生成を実現する。

3. 手法

ここではマスク画像から画像生成する Image-to-image 変換のネットワークにスタイル特徴を抽出し、スタイルを考慮した画像生成について説明する。まず、スタイル画像から各要素のスタイルを抽出する Style encoder について説明し、次に Image-to-image 変換のネットワークである SPADE の手法と作成した Style encoder に組み合わせについて説明する。

3.1 Style Encoder

スケッチ画像を基にして生成した画像のスタイルを制御するため，与えられたスタイル画像から領域マスク要素のマスクスタイル特徴を抽出する Style Encoder を提案する。Style Encoder は semantic segmentation mask とスタイル画像を受け取り、スタイル画像の各マスク要素（顔の場合は髪の毛、皮膚、口の領域）からスタイル特徴を抽出する。

図 1 の上段は、Encoder-Decoder スタイル特徴抽出器と mask pooling レイヤーからなる mask style encoder の基本的な構成を示している。まず、Convolution レイヤーと Transposed convolution レイヤーからなる Encoder-Decoder ネットワークにスタイル画像からスタイル特徴マップを抽出する。 次に、スタイル画像に対応する seman-

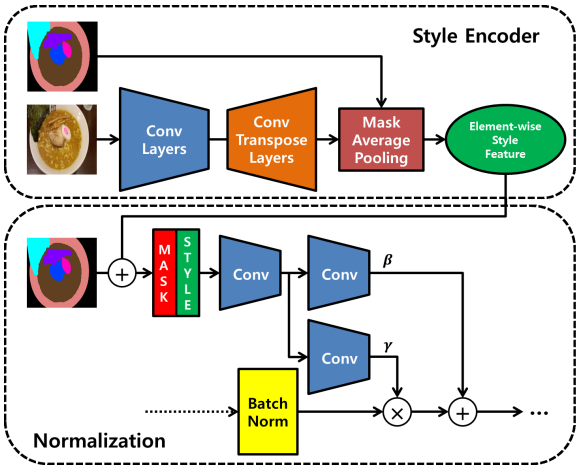

図 1 Style Encoder と Spade Normalization を組み合わせた構造図。

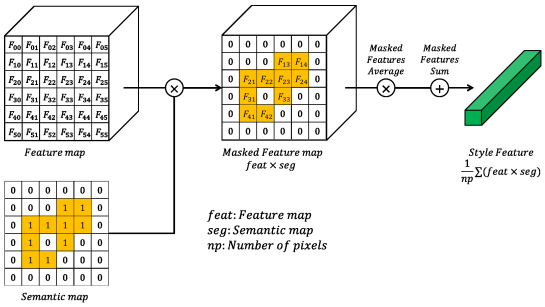

図 2 Mask Average Pooling の流れ図。

tic segmentation mask を補助入力とする Mask Average Pooling を特徴マップに適用し、各マスク要素のスタイル特徴を抽出する。

図 2 に示すように、入力された特徴マップの特徴ベクトルを特定の領域ラベル上で平均化した特徴ベクトルを抽出する Mask Average Pooling レイヤーを提案する。特徴マップをチャネル方向に並んだ特徴ベクトルの集合とみなす。 図は mask average pooling レイヤーが対応する空間的位置（黄色で強調表示）の特徴ベクトルを特定領域ラベルに対して平均化していることを示している。 この計算を全てのマスクラベルに対して繰り返すことで、図に示すラーメン画像の場合には、背景領域、器領域、スープ領域などの segmentation mask 領域要素の平均化されたスタイル特徴量を抽出することができる。これらのマスク要素に対応するマスクスタイルベクトルを用いて、マスクスタイルベースの画像合成を行う。なお、segmentation mask は mask average pooling レイヤーの入力特徴マップと同じサイズにリサイズされる。

3.2 損失関数

本研究の損失関数は Pix2pixHD[10] と SPADE[8] の 3 つの損失関数に Style encoder を追加して抽出したスタイルを反映するように改良した損失関数を使用した。

(1) Adversarial 損失

Conditional Adversarial 損失は式 1 のミニマックスゲームを通して、実画像の条件付き分布をモデリングする。G は Generator、D_1 と D_2 は Discriminator、E は Style encoder である。

$$\mathcal{L}_{adv}(G,D,E) = \min_{G,E}\max_{D_1,D_2}\sum_{k=1,2}\mathcal{L}_{GAN}(G,D_k,E) \tag{1}$$

$\mathcal{L}_{GAN}$ は次の式 2 のような Hinge 損失関数を用いている。その式で x は入力画像、s は x の領域分割したマスク画像である。

$$\begin{aligned}\mathcal{L}_{GAN}(G,D,E) = &-\mathbb{E}[min(0,-1+D_k(s,x))] \\ &-\mathbb{E}[min(0,-1-D_k(s,G(s,E(s,x))))]\end{aligned} \tag{2}$$

(2) Feature matching 損失

Discriminator D_k の i 番目レイヤーの特徴マップを $D_k^{(i)}$ と定義する。また、T は D_k のすべてのレイヤーの数、N_i は各レイヤーでの要素の数である。

$$\begin{aligned}\mathcal{L}_{FM}(G,D_k,E) =&\mathbb{E}\sum_{i=1}^{T}\frac{1}{N_i}[\|D_k^{(i)}(s,x) \\ &-D_k^{(i)}(s,G(s,E(s,x)))\|_1]\end{aligned} \tag{3}$$

(3) Perceptual 損失

$F^{(i)}$ は VGG ネットワークの i 番目のレイヤーであり、M_i は VGG ネットワークの要素の数である。

$$\begin{aligned}\mathcal{L}_{perceptual}(G,F,E) =&\mathbb{E}\sum_{i=1}^{N}\frac{1}{M_i}[\|F^{(i)}(x) \\ &-F^{(i)}(G(s,E(s,x)))\|_1]\end{aligned} \tag{4}$$

最後に、その 3 つの損失関数を組み合わせて画像生成の学習を行う。本研究では、$\lambda_1 = \lambda_2 = 10$ として学習を行った。

$$\begin{aligned}\mathcal{L}_{total} = \min_{G,E}\Bigg(&\Big(\max_{D_1,D_2}\sum_{k=1,2}\mathcal{L}_{GAN}(G,D_k,E)\Big) \\ &+\lambda_1\sum_{k=1,2}\mathcal{L}_{FM}(G,D_k,E)+\lambda_2\mathcal{L}_{perceptual}(G,F,E)\Bigg)\end{aligned} \tag{5}$$

4. 実験

4.1 データセット

本研究では、555 枚のラーメン画像と各要素ごとにラベルが付けられたマスク画像で構成されているラーメン画像データセット UEC-Ramen555[1] を使用した。また、このデータセットのマスク画像はマスク画像の各要素は 15 クラスにカテゴリで構成している。データセットの例は図 3 である。

本実験では、データセットの中で 500 枚ずつの実画像とマスク画像を用いて各タスクの学習を行い、55 枚ずつの実画像とマスク画像を使用してテストを行う。また、スープカテゴリの明確な確認のために、カテゴリの中で塩スープ、醤油スープ、味噌スープ、豚骨スープ、辛口スープの 5 つのスープカテゴリを 1 つのスープカテゴリに統合して、画像生成を行った時にスープのスタイルが良く反映されたかも確認した。

図 3 ラーメン画像のデータセットの例 （上：元画像、下：各要素を領域分割したマスク画像）。

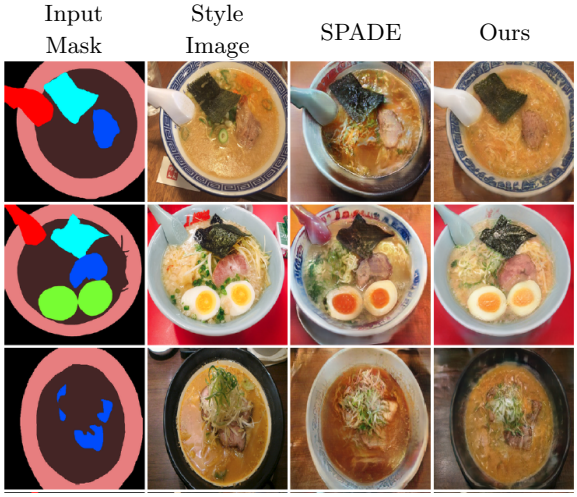

図 4 スタイルを考慮した画像生成結果の比較 （1 列目：入力マスク画像、2 列目：入力スタイル画像、3 列目：SPADE の結果、4 列目：本研究の手法の結果）

表 1 スタイルを考慮した画像生成結果を FID とユーザー評価を実施した結果

	FID	ユーザー評価
SPADE	**68.91**	8.67%
Ours	73.28	**91.33%**

4.2 スタイルを考慮した画像生成

本実験では、まず、スタイル画像から抽出したスタイルとスタイル画像に対応するマスク画像を用いた画像生成を行って、従来手法と比較して評価を行う。次に、1 つのスタイル画像から抽出したスタイルと任意のマスク画像を用いた画像生成を行う。最後に、ラーメン以外のデータセットに適用した例も示す。

最初に、1 つのスタイル画像とそのスタイル画像に対応するマスク画像を用いてスタイルを考慮した画像生成を行い、従来手法である SPADE と生成結果の比較を行った。生成結果は図 4 で表せる。結果の図としては、1 列目は入力マスク画像、2 列目は入力スタイル画像、3 列目は生成した結果画像である。評価には、Fréchet Inception Distance(FID) とユーザー評価を用いて各手法を評価を行った。FID を用いた評価では, 生成画像とラーメンデータセットの画像 555 枚を用いて評価を行った。ユーザー評価では、従来手法と提案手法から生成された結果でスタイル画像からのスタイルが良く反映された手法の結果を選択することで評価を行った。評価の結果は表 1 に示す。

従来手法である SPADE と本研究の手法の結果を比較すると、FID を用いた評価は従来の手法が本研究の手法より良いスコアが出る結果が得られ、従来手法がより多様なラーメン画像の生成が可能であることをわかった。また、従来手法は特定のスタイルを指定して生成する訳ではないので、生成しやすいスタイルで画像生成を行う傾向がある。

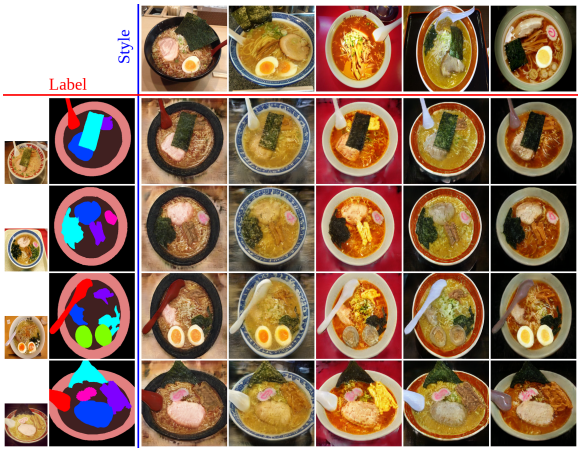

図 5　特定のスタイル画像から抽出したスタイルを考慮した画像生成の結果。

そのため、FID を用いた評価は従来手法のほうが良い結果になったと考えられる。しかしながら、入力したスタイルが良く反映されたかどうかを結果画像とユーザー評価の結果で比較すると、本研究の手法の方が既存手法よりカテゴリ要素のスタイルを良く反映させて画像生成を行ったことが見られる。

次に、特定のスタイル画像からスタイル特徴を抽出し、任意のマスク画像と抽出したスタイルを基にして、スタイルを考慮した画像生成を行う。図 5 は実験結果として、1 行目のスタイル画像からスタイル特徴を抽出したスタイルと 1 列目のマスク画像を用いてスタイルを考慮した画像生成を行った結果である。この結果から、複数の特定画像から必要なスタイル特徴だけを抽出し、その様々なスタイル特徴の組み合わせを用いた画像生成が可能であることが分かる。

次に、複数のスタイル画像から特定のスタイルを抽出し、マスク画像と抽出したスタイルを用いて画像生成を行った。また、本実験で抽出するスタイルは、結果の明確な比較のために、3 つのカテゴリ要素（背景、器、スープ）に限って、スタイル画像からスタイルを抽出し、スタイルを考慮した画像生成を行う。結果を図 6 に示す。1 行目はスタイル画像と各スタイル画像から抽出したスタイルのカテゴリがマスク画像で示されている。ここでは、スペースの都合で 2 枚の画像からのみスタイルを抽出しているが、実際にはトッピングの領域インスタンスごとに異なるスタイルを指定することが可能で、厳選されたトッピングのスタイルを集めて、「究極の一杯」を生成することが可能である。なお、提案手法はラーメン以外の通常のマスク画像が付随した画像データセットにも適用可能で、CelebAMask-HQ に適用した例を図の右に示す。

5. おわりに

本研究では、Image-to-image 変換のネットワークと Style encoder を用いてスタイルを考慮したラーメン画像生成を行った。従来手法では各要素ごとのスタイルの制御ができないので、入力画像と生成画像の要素の色などのスタイルが異なる問題があった。そのため、入力したスタイル画像のスタイルを考慮する画像生成を行うために、Style encoder を作成し、SPADE を改良することで、スタイルを考慮した画像生成するネットワークを作成した。実験でラーメン画像データセットを用いて、ユーザーが簡単にスケッチした画像と好みに合わせたスタイルを持っているラーメン画像を用いて新しいラーメン画像を生成した。また、複数の画像からそれぞれ希望するスタイル特徴を抽出して、スタイルを反映した画像生成も可能だった。今後の研究では、より詳細なスタイル特徴を抽出し、より詳細なスタイル制御画像合成を行う予定である。

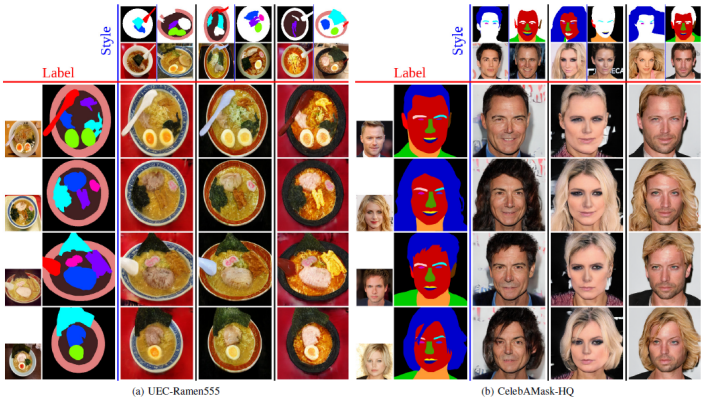

図 6　複数のスタイル画像から抽出したスタイルを考慮した画像生成の結果。左は CelebAMask-HQ による結果。

謝辞　本研究は JSPS 科研費 15H05915, 17H01745, 19H04929, 17H06100 の助成を受けたものです.

参考文献

[1] Cho, J., Shimoda, W. and Yanai, K.: RamenAsYouLike: Sketch-based Food Image Generation and Editing, *ACM Multimedia, Demo Track,* (2019).

[2] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S. and Choo, J.: StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation, *CVPR* (2018).

[3] Huang, X., Liu, M. Y., Belongie, S. and Kautz, J.: Multimodal Unsupervised Image-to-image Translation, *ECCV* (2018).

[4] Isola, P., Zhu, J. Y., Zhou, T. and Efros, A. A.: Image-To-Image Translation With Conditional Adversarial Networks, *CVPR* (2017).

[5] Lee, H. Y., Tseng, H. Y., Huang, J. B., Singh, M. and Yang, M. H.: Diverse Image-to-Image Translation via Disentangled Representations, *ECCV* (2018).

[6] Liu, M. Y., Breuel, T. and Kautz, J.: Unsupervised Image-to-Image Translation Networks, *NeurIPS* (2017).

[7] Mirza, M. and Osindero, S.: Conditional Generative Adversarial Nets, *arXiv:1411.1784* (2014).

[8] Park, T., Liu, M. Y., Wang, T. C. and Zhu, J. Y.: Semantic Image Synthesis With Spatially-Adaptive Normalization, *CVPR* (2019).

[9] Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, *MICCAI* (2015).

[10] Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J. and Catanzaro, B.: High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs, *CVPR* (2018).

[11] Zhu, J. Y., Park, T., Isola, P. and Efros, A. A.: Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks, *ICCV* (2017).

[12] Zhu, P., Abdal, R., Qin, Y. and Wonka, P.: SEAN: Image Synthesis with Semantic Region-Adaptive Normalization, *CVPR* (2020).