

Conditional GAN を用いたスケッチからのフォトリアルな画像生成

落合 晃汰[†] 青野 雅樹^{††}

[†] 豊橋技術科学大学 博士前期課程情報・知能工学専攻 〒441-8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1

^{††} 豊橋技術科学大学 情報・知能工学系 〒441-8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1

E-mail: [†]ochiai.kota.yx@tut.jp^{††}aono@tut.jp

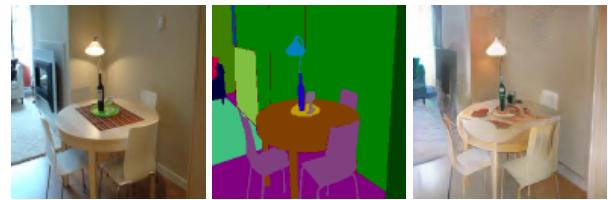
あらまし 本研究では, pix2pixHD をベースとした Conditional GAN を使用して, 線と色からなるスケッチからフォトリアルな画像を生成する手法を提案する。Conditional GAN はセグメンテーションマップやバウンディングボックスからの画像生成を可能にしたが, これらの手法ではラベルに依存するために汎化性能が低く, 広く実用化されるまでには至っていない。そこで, 本研究ではラベルを用いない手法として物体の輪郭線と色からなるスケッチ画像を新たな入力形式として提案する。さらに, 生成モデルとして Fast Fourier Convolution と Patchwise Contrastive Loss を組み合わせた手法を提案し, FID 評価尺度で精度改善できたことを報告する。

キーワード 深層学習, GAN, 画像生成

1 はじめに

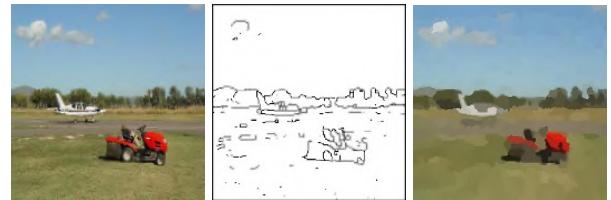
近年, 3DCG の進化とともにフォトリアルな CG が多く使用されるようになった。映画やアニメなどの映像作品以外にも, 建築, 工業製品などの表現によく用いられる。このような 3DCG は物体の形状, 質感, 光の移動を明示的にシミュレーションすることができるため, 非常に美しくかつ正確な画像を生成することができる。しかし, 高度なスキル及び膨大な作業時間を要するというデメリットも存在する。そこで, 特別なスキルを用いずに短時間で画像生成を可能とする手法として, 図 1 に示すような, セグメンテーションマップからの画像生成が注目されている。セグメンテーションマップは画像のピクセル一つ一つに対して, 何が写っているかと言った, ラベルやカテゴリを対応付けた画像である。この方法は, ラベルやカテゴリ情報を画像に指定するだけで良いため非常に簡単に画像生成が可能となるという利点がある。一方で欠点も存在し, ラベルにある物体しか生成することができないという問題がある。また, ラベルをどのような基準でどこまで分類すればいいかという点も問題となる。例えば, 料理の画像を生成したいときに「食べ物」というラベルしか存在しない場合, 希望とする画像が生成されるとは限らない。しかし, 数多あるすべての料理画像に対して教師用のセグメンテーションマップを用意することは困難である。

これらを解消するためのラベルに依存しない方法として, 図 2 に示すような, 物体の輪郭線と色からなるスケッチ画像を新たな入力形式として提案する。スケッチからのフォトリアルな画像の生成は, セグメンテーションマップからの画像生成と同様に, pix2pix の手法で実現できる。これは敵対的生成ネットワーク (Generative Adversarial Network:GAN) を利用した画像変換手法の一種である。GAN は生成ネットワークと識別ネットワークの 2 つのネットワークから構成される。生成ネットワークは文字通り目的とする生成物を生成するためのネットワークで, 識別ネットワークは生成ネットワークが生成した生成物と



(a) Ground truth (b) Input (c) Output

図 1: セグメンテーションマップからの画像生成. (a) は実際の写真, (b) は (a) に対応するセグメンテーションマップ. 生成モデルに (b) を入力することで (c) の画像が得られる.



(a) Ground truth (b) Edge (c) Color



(d) Input (e) Output

図 2: スケッチからの画像生成. (a) は実際の写真, (b) は (a) に対応する物体の輪郭線画像, (c) は (a) に対応する色画像. (d) は (b) と (c) を乗算合成することにより得られる, (a) に対応するスケッチ. 生成モデルに (d) を入力することで (e) が得られる.

本物を識別するためのネットワークとなっている。この 2 つのネットワークを互いに競い合わせることで本物の特徴を捉えた生成ができるようになる。本研究では, pix2pix から派生した

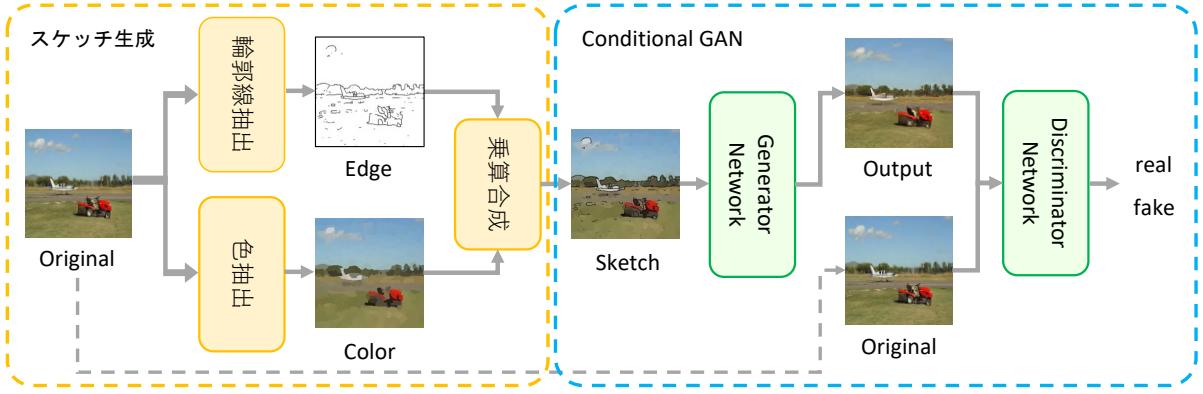


図 3: 提案手法の流れ. スケッチ生成部により, 写真に対応する人が描いたようなスケッチを自動生成する. Conditional GAN 部ではスケッチから写真の状態に復元することを目的として学習を行う.

pix2pixHD をベースとして Fourier Convolution と Patchwise Contrastive Loss を組み合わせた手法を提案し, 精度を改善したことを示す.

2 関連研究

近年の画像生成は, Conditional GAN を用いた手法が普及しており, 研究が活発に行われている. また, 画像生成以外にも画像変換や画像修復に対してもよく用いられる. 以下では, Conditional GAN をベースとする画像生成や画像変換, 画像修復に関する研究について述べる.

2.1 画像生成・変換

GAN [1] から派生した手法として Conditional GAN [2] が存在する. 通常の GAN はノイズベクトルのみを入力とするが, Conditional GAN は条件ベクトルを入力とする. これによって, 条件ベクトルに対応した生成ができる. Isola らの pix2pix [3] はその条件として画像を用いる手法である. pix2pix は, 通常の Conditional GAN の損失関数に加えて Discriminator の損失関数に L1 損失関数が加えられ, さらに画像を小さなパッチに分割し, パッチ単位で本物か偽物かの判別を行っている. これにより画像の全体像を L1 損失関数で捉え, 詳細な部分は Conditional GAN で捉えられるようになり, 精度の高い画像変換が可能となっている. また, Generator に対して U-Net 構造を導入する工夫を行い, さらに精度を向上させている. pix2pix は画像のペアがあればその対応関係を学習することができるため, 線画の写真化やセグメンテーションマップからの画像生成, セマンティックセグメンテーション, モノクロ画像のカラー化, 航空写真から地図を生成するなど様々なタスクに使用することができる.しかし, 256×256 の低解像度なものまでしか扱うことができないという問題がある.

これに対して Wang らは, 高解像度な画像変換を可能にするモデルとして pix2pixHD [4] を提案した. pix2pixHD は, 高解像度画像を効率よく判別するために異なるスケールの複数の Discriminator を用意して小さなスケールに変換した画像に対しても Discriminator を適応できるように工夫されている. こ

れにより pix2pixHD は, 2048×1024 までのより高解像度の画像生成を可能とした. さらに, 通常の損失関数に加えて, Feature Matching Loss を定義している. これは, Discriminater に Ground Truth 画像と生成画像を入力した際で各層の出力を一致させるためのものであり, これにより Generator により自然な生成を促すことができる.

pix2pix 及び pix2pixHD は対になっている画像の変換を対象とする手法であったが, Taesung らは, 対になつてない画像変換を可能にする CUT(Contrastive Learning for Unpaired Image-to-Image Translation) [5] を提案した. この手法以前は, 対になつてない画像変換を行うために CycleGAN [6] を使用する必要があった. CycleGAN は双方向に変換を行うため, Generator を 2 つ用意する必要がある. これにより, 学習時間が比較的長く, メモリ効率も悪いという問題があったが, CUT では Patchwise Contrastive Loss を導入することにより, 1 つの Generator で学習することができる. これにより学習時間とメモリ使用量を抑え, 精度も向上させることができる.

2.2 画像修復

画像修復は, 画像の削除された領域を違和感のない形で埋める作業である. Jiahui らが提案した Deepfillv1 [7] は, 画像の一部が長方形に切り取られた画像の修復が可能なモデルであり, 従来から性能を大きく向上させた. しかし, 修復領域のエッジの修復が上手く行かず, 修復領域が浮いて見えてしまう問題があった. これに対して, 強化されたモデルである Deepfillv2 [8] は, 長方形に切り取られた画像だけでなく, フリーフォーム切り取り領域に対しての修復が可能となった. さらに, Gated Convolution Layer を導入し, 違和感の少ない修復が可能となった.

Suvorov らは, より高画質な画像に対しても高精度な画像の修復を実現するモデルとして LaMa [9] を提案した. LaMa は FFT(Fast Fourier Transform) [10] を使った畳み込みである FFC(Fast Fourier Convolution) [11] を導入したモデルであり, より広い受容野を獲得することで精度を向上させている. さらに, 256×256 の低解像度なデータセットで学習して高解像度で推論しても性能の低下が起りづらいという特徴もある.

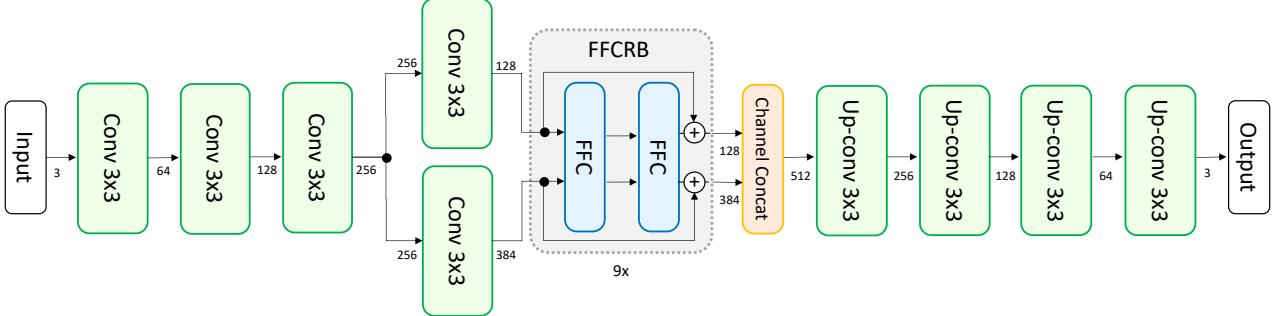


図 4: FFC(Fast Fourier Convolution) で構成された Residual Block をボトルネックとする Generator

3 提案手法

提案手法の流れを図 3 に示す。ユーザの入力は「描写したい物体の輪郭線を描いた線画」と「対応する色が塗られた色画像」である。この線画と色画像は人が描くことを想定しているが、本研究では、この線画と色画像を Ground Truth 画像に対して画像処理を行うことで自動的に抽出する。さらに、得られた線画と色画像を乗算合成することで 1 枚のスケッチ画像とし、それを入力として、Conditional GAN の学習を行う。以下では、スケッチの生成方法及び提案する生成モデルについて説明する。

3.1 スケッチの生成

線の抽出方法としてはノイズ除去を施し、グレースケール画像と白部分を膨張させたグレースケール画像の差を取り白黒反転させ、輪郭線を抽出する。さらに、その画像に対して Zhang-Suen アルゴリズム [12] を適応し細線化を行うことで、よりはっきりとした線に加工する。

色の抽出方法は、スーパーピクセルのアルゴリズムの一つである SEEDS(Superpixels Extracted via Energy-Driven Sampling) [13] を用いて抽出を行う。スーパーピクセルとは色が類似する領域をグルーピングするアルゴリズムで各領域を平均した色で塗りつぶしたものを作成する。

3.2 Conditional GAN

生成手法は pix2pixHD の Generator 部と損失関数部に変更を加えたものである。具体的には FFC で構成された Residual Block をボトルネックとする Generator に変更し、損失関数として Patchwise Contrastive Loss を導入する。それ以外の Discriminator 等の構造については pix2pixHD と同様である。以下では FFC と Patchwise Contrastive Loss を説明する。

3.2.1 FFC(Fast Fourier Convolution)

2.2 節で述べたように FFC(Fast Fourier Convolution) とは画像全体の特徴量を使用するための、FFT(Fast Fourier Transform) を使った畳み込みである。FFC ではチャネルをローカルチャネルとグローバルチャネルに分割する。ローカルチャネルでは従来の畳み込みを行い、グローバルチャネルでは FFT を利用した畳み込みを行う。FFT での畳み込みは FFT した結果の周波数領域の画像に通常の畳み込みを行い、空間構造を復元するためにその結果に逆 FFT を行うことで実現する。これに

より、画像全体をカバーする畳み込みが可能となる。本研究では、LaMa [9] と同様にローカルチャネル数は 128、グローバルチャネル数は 384 としている。図 4 に Generator のネットワーク図を示す。

3.2.2 Patchwise Contrastive Loss

Patchwise Contrastive Loss とは教師なし画像変換において Contrastive learning を活用する損失関数である。2.2 節で述べたように、1 つの Generator で教師なし画像変換を可能にするための損失関数であるが、教師あり学習であるセグメンテーションマップからの画像生成においても効果的であることが示されている [5]。Patchwise Contrastive Loss は画像全体ではなく、パッチ単位かつ Generator の多層で Contrastive Learning を行う。また、パッチのネガティブサンプルは他の画像ではなく、入力画像内からサンプリングする。

4 実験

4.1 データセット

本研究では、ADE20K データセット [14] と GTA5 データセット [15] の 2 つを使用する。

4.1.1 ADE20K データセット

ADE20K データセットは屋内および屋外の様々なシーンを含む画像データセットである。データセットは 20,210 枚の学習用画像と 2,000 枚の検証用画像から構成される。画像は学習用と検証用共に 256×256 にリサイズし、実験を行った。

4.1.2 GTA5 データセット

GTA5 データセットはオープンワールドゲーム Grand Theft Auto V で、車からの視点の街並みや道路の画像から構成されるデータセットである。データセットは 12,403 枚の学習用画像と 6,382 枚のテスト画像から構成される。画像サイズは学習用と検証用共に 256×512 である。

4.2 ベースラインと提案手法

提案手法として、FFC と Patchwise Contrastive Loss を導入した pix2pixHD に加えて、FFC のみを導入した pix2pixHD, Patchwise Contrastive Loss のみを導入した pix2pixHD, U-Net [16] を導入した pix2pixHD, FFC と U-Net を導入した pix2pixHD, Patchwise Contrastive Loss と U-Net を導入した pix2pixHD, FFC と U-Net と Patchwise Contrastive Loss を導入した pix2pixHD の 7 つを提案する。

表 1: ADE20K データセットでの実験結果と各モデルの計算量. FID スコアは生成結果と Ground truth 間の特徴距離を表し, FLOPs は Generator ネットワークに 256×256 の RGB 画像を入力した際の浮動小数点演算数を表す. FID スコアについては, Ground truth 画像と入力画像であるスケッチ間の値が ADE20K データセットでは 134.99 でありこの値をどれだけ小さくできるかが精度の目安となる.

System	FID	FLOPs(G)
pix2pix	51.65	36.31
deepfillv2	32.75	274.30
deepfillv2 (U-Net)	26.73	495.08
pix2pixHD	32.84	125.86
pix2pixHD (FFC)	33.21	85.3
pix2pixHD (PCL)	26.02	125.86
pix2pixHD (U-Net)	26.02	156.09
pix2pixHD (FFC, PCL)	26.51	85.3
pix2pixHD (FFC, U-Net)	32.05	115.53
pix2pixHD (PCL, U-Net)	27.10	156.09
pix2pixHD (FFC, PCL, U-Net)	26.61	115.53

表 2: GTA5 データセットでの実験結果. FID スコアは生成結果と Ground truth 間の特徴距離を表す. Ground truth 画像と入力画像であるスケッチ間の値が GTA5 データセットでは 196.43 でありこの値をどれだけ小さくできるかが精度の目安となる.

System	FID
pix2pix	37.73
pix2pixHD	19.85
pix2pixHD (FFC, PCL)	15.28

これらをベースライン手法である, pix2pix [3], pix2pixHD [4] , deepfillv2 [8], U-Net を導入した deepfillv2 の 4 つの手法と比較する. GTA5 データセットについては, FFC と Patchwise Contrastive Loss を導入した pix2pixHD と pix2pix と pix2pixHD の 3 つの手法と比較する.

4.3 評価指標

評価指標には Fréchet Inception Distance(FID) [17] を使用して, 生成結果の分布と Ground truth の分布の間の特徴距離を測定する. また, flops-counter.pytorch [18] を使用して, Generator ネットワークに 256×256 の RGB 画像を入力した際の FLOPs(Floating-point Operations) を計算する. これにより, 各モデルの計算量である浮動小数点演算数を求める.

4.4 モデルパラメータ

epoch 数は 50, バッチサイズは 4 とした. その他のパラメータは pix2pix はオリジナルと同様の値, pix2pixHD をベースとする手法は pix2pixHD と同様の値, deepfillv2 をベースとする手法は deepfillv2 と同様の値で実験を行った.

4.5 実験結果

表 1 は ADE20K データセットでの実験結果と各モデルの計算量であり, 図 5 はそれをグラフにプロットしたものである. FID スコアについては, Ground truth 画像と入力画像である

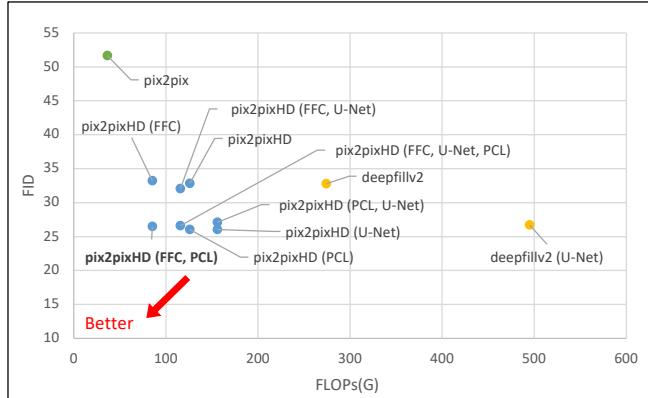


図 5: ADE20K データセットでの FID スコアと各モデルの FLOPs. 横軸が FLOPs で縦軸が FID スコアであり, 左下へ行くほど軽量かつ高精度であることを表す.

スケッチ間の値が ADE20K データセットでは 134.99 であり, GTA5 データセットにでは 196.43 であった. この値をどれだけ小さくできるかが精度の目安となる. 表 1 より提案手法のなかで pix2pixHD に Patchwise Contrastive Loss を導入した手法と U-Net を導入した手法がそれぞれ最も高い精度となつた. pix2pixHD と deepfillv2 に関しては U-Net を導入することで精度が向上したが, FFC と Patchwise Contrastive Loss を導入した pix2pixHD に関しては U-Net を導入した場合でも精度の向上は見られなかった. また図 5 より, FFC と Patchwise Contrastive Loss を導入した pix2pixHD は他の高精度なモデルと同等の精度でありながら, pix2pix に次いで Generator の FLOPs が小さく, 計算量が少ないことが読み取れる.

図 6 は ADE20K データセットでのベースライン手法および提案手法の 1 つである FFC と Patchwise Contrastive Loss を導入した pix2pixHD の生成例で, 図 7 は GTA5 データセットでの生成例である. 図 6 よりベースライン手法では物体の輪郭がぼやけている上にノイズのようなものが入っていることが確認できるが, 提案手法ではそれが軽減されていることが確認できる. 特に, 1 列目と 2 列目の例について木や葉の質感が上手く表現されていることがわかる. しかし, 5 列目と 6 列目の例については人の顔が復元できていないことも読み取れる. 図 7 の GTA5 データセットでの生成例については, pix2pix ではノイズや白飛びが確認でき, pix2pixHD でも白飛びや輪郭のぼやけが確認できる. これに対して提案手法では全体的に物体の輪郭がはっきりとし, 余計なノイズも軽減されていることが確認できる. さらに, アスファルトの質感については Ground truth と遜色ない品質で復元できていることが確認できる. また, ADE20K の生成例と同様に木や芝生の質感が上手く表現されていることが分かる.

4.6 考察

実験結果より, 提案手法の 1 つである FFC と Patchwise Contrastive Loss を導入した pix2pixHD は, 他の高精度なモデルと同等の精度を保ちながら軽量化することに成功した. FFC については, フーリエ変換により広域の特徴量を得ることができ

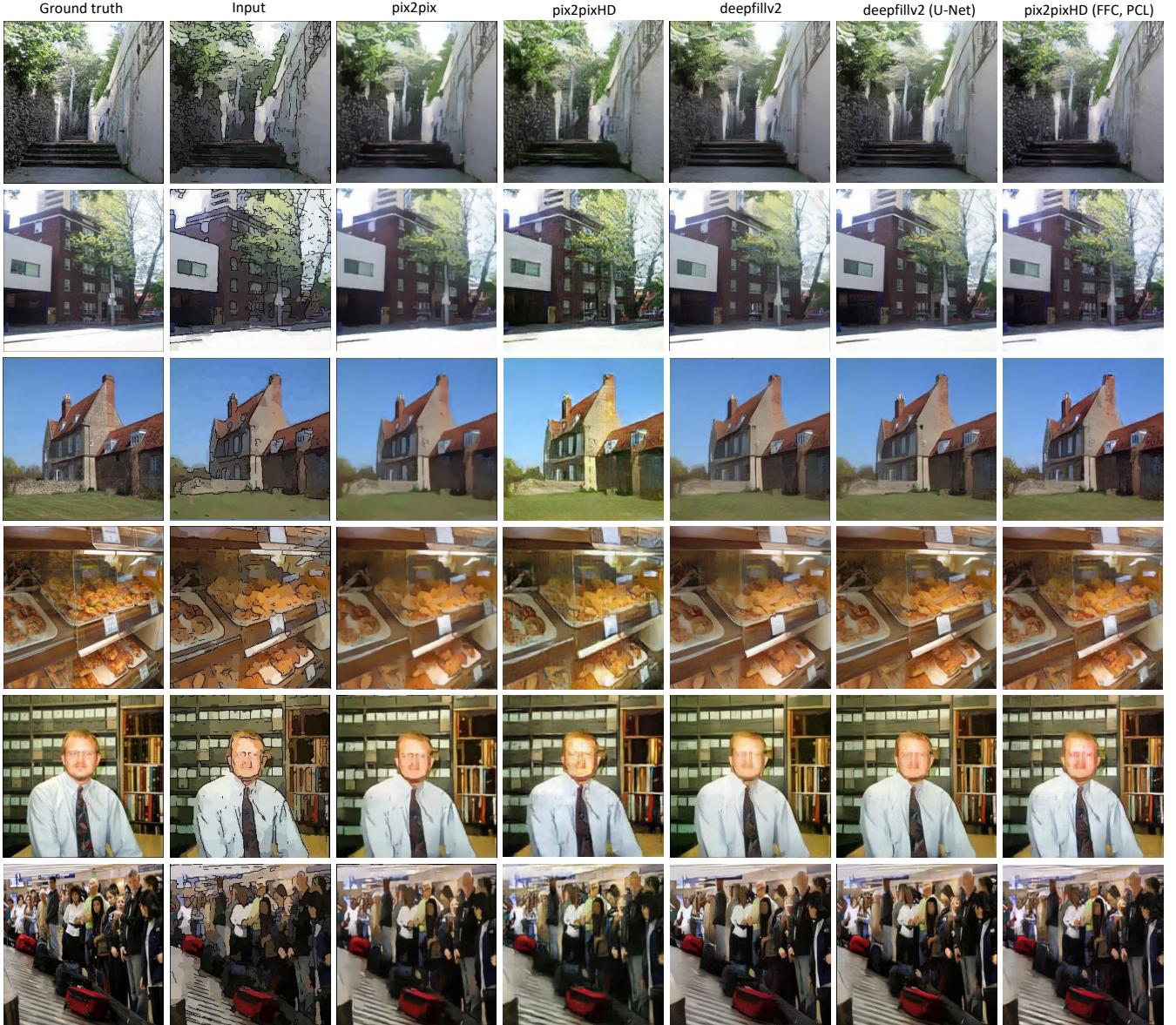


図 6: ADE20K データセットでの各手法の生成例. Ground truth は実際の写真, Input が入力するスケッチ, pix2pix, pix2pixHD, deepfillv2, deepfillv2(U-Net) はベースライン手法, pix2pixHD(FFC, PCL) は提案手法である.

るため, 画像変換をする上でスケッチからより多くの手がかりを得ることに繋がり, 少ない計算量で精度が向上したと考えられる. Patchwise Contrastive Loss については, 入力であるスケッチ画像と出力画像における同じパッチの領域が似るように学習するため, スケッチの構造を保持しつつフォトリアルに変換することを補助したと考えられる.

また生成結果では, 全体的に ADE20K よりも GTA5 データセットの結果のほうが物体の輪郭がはっきりとし生成品質が高い. さらに, FID スコアでも pix2pix, pix2pixHD, 提案手法の GTA5 データセットで実験したすべての手法において ADE20K データセットでの結果よりも精度が良い結果となった. これは, ADE20K データセットが様々なシーンを含むデータセットであることに対して, GTA5 データセットが道路や街並みの画像のみのデータセットであり, ADE20K データセットと比べると変換が容易であったためだと考えられる.

5 おわりに

本研究では, ラベルに依存しない画像生成の入力形式として, 物体の輪郭線と色からなるスケッチ画像を提案するとともに, その生成手法として FFC と Patchwise Contrastive Loss を導入した pix2pixHD を提案した. 実験結果より, FFC によって大きな受容野を得ることはスケッチからの画像生成においても精度の向上に効果的であることが分かった. しかし, FFC だけではなく Dilated Convolution [19] や Vision Transformer [20] も大きな受容野を得る方法として有効な選択肢であることが示されている [9]. そのため, これらの手法を導入した場合についても検討する必要があると考えられる.

また, Patchwise Contrastive Loss を導入することで精度向上も確認したが, 本来は対になつてない画像データセットでの変換を目的として考案された損失関数であるため, 対になつ



図 7: GTA5 データセットでの各手法の生成例. Ground truth は実際の写真, Input が入力するスケッチ, pix2pix, pix2pixHD, はベースライン手法, pix2pixHD(FFC, PCL) は提案手法である.

ている場合の画像変換を考慮した構造に変更することで精度の向上が可能であると考えられる. また、ラベルに依存しない入力形式としてスケッチ画像を提案したが、現在の入力画像は Ground truth 画像の情報を多く保持しており、入力画像の時点である程度フォトリアルであると言える. Ground truth 画像と比較すると粗いとはいえ、ここまで正確なスケッチを人が描くためには、ある程度高度なスキルを必要とすると考えられる。

今後の課題としては、スケッチとセマンティックセグメンテーションの両方を入力可能にし、ラベルへの依存を軽減しつつ入力を簡単にすることが挙げられる。

文 献

- [1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv preprint arXiv:1406.2661.
- [2] Mehdi Mirza, and Simon Osindero. 2014. Conditional Generative Adversarial Nets. arXiv preprint arXiv:1411.1784.
- [3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. arXiv preprint arXiv:1611.07004.
- [4] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2017. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. arXiv preprint arXiv:1711.11585.
- [5] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. 2020. Contrastive Learning for Unpaired Image-to-Image Translation. arXiv preprint arXiv:2007.15651.
- [6] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. arXiv preprint arXiv:1703.10593.
- [7] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative Image Inpainting with Contextual Attention. arXiv preprint arXiv:1801.07892.
- [8] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. 2019. Free-Form Image Inpainting with Gated Convolution. arXiv preprint arXiv:1806.03589.
- [9] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv preprint arXiv:2109.07161.
- [10] G. D. Bergland. 1969. A guided tour of the fast fourier transform. IEEE Spectrum, vol.6(7), 41–52.
- [11] Lu Chi, Borui Jiang, and Yadong Mu. 2020. Fast fourier convolution. Advances in Neural Information Processing Systems, vol.33, 4479–4488.
- [12] T. Y. Zhang, and C. Y. Suen. 1984. A fast parallel algorithm for thinning digital patterns. Communications of the ACM, vol.27, no.3, 236–239.
- [13] Michael Van den Bergh, Xavier Boix, Gemma Roig, and Luc Van Gool. 2013. SEEDS: Superpixels Extracted via Energy-Driven Sampling. arXiv preprint arXiv:1309.3848.
- [14] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. CVPR 2017. 5122–5130.
- [15] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for Data: Ground Truth from Computer Games. arXiv preprint arXiv:1608.02192.
- [16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015.

- U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv preprint arXiv:1505.04597.
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv preprint arXiv:1706.08500.
- [18] sovrasov. 2021. flops-counter.pytorch.
<https://github.com/sovrasov/flops-counter.pytorch>. (2022).
- [19] Fisher Yu, and Vladlen Koltun. 2015. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv preprint arXiv:1511.07122.
- [20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.