

画像変換 AI における画像の情報量に依る生成画像精度の非対称性について

A study on the asymmetry of the generated image accuracy depending on the entropy of the image in image translation AI

松川 瞬* 真田 博文* 稲垣 潤*

Shun MATSUKAWA, Hirofumi SANADA, Jun INAGAKI

Abstract

In this study, we demonstrated the asymmetry between the concretization task and the abstraction task by pix2pix, which is the basis of current image transformation AI.

For the AtoB model, which converts satellite images to map images, and the BtoA model, which converts map images to satellite images, the entropy of each image was calculated after obtaining the GLCM of each image, and the correlation between the entropy ratio of input image to output image and the accuracy of the generated image by pix2pix was obtained. The results showed that there is a negative correlation in the AtoB model, which is responsible for the abstraction task, where the accuracy decreases as the image entropy ratio increases, and only a weak negative correlation in the BtoA model, which is responsible for the concretization task. Then it has been demonstrated that pix2pix have an asymmetric property between abstract tasks than concrete tasks.

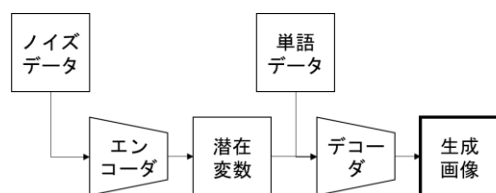
We investigated this property by focusing on the L1 error of U-Net, which is frequently used in image generation models, and found that the convergence speed in the AtoB model is faster than in the BtoA model. From this, we infer that the solution space of the weights is sparser in the abstraction task, and that learning proceeds toward larger amounts of entropy due in part to the skip-connection effect of the U-Net structure. In the future, as the characteristics of U-Net become clearer, it can be used to improve the accuracy of image abstraction and lower dimensionality.

1. はじめに

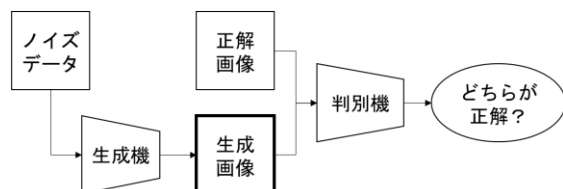
ここ 1～2 年で, Midjourney⁽¹⁾ や stable diffusion⁽²⁾, NovelAI⁽³⁾ といった, まるで実写か人間が描いたかのような画像を生成できる AI モデルが現れはじめている。それらは, ガウスノイズ画像に加え, 生成したい画像の概念を表現する単語列を入力に含め, ノイズ画像を単語列の意味に沿った画像に変換し出力する (すなわち生成する) ネットワーク構造になっており, 拡散モデル⁽⁴⁾ と呼ばれる (図 1(a))。以前から用いられていた敵対的生成ネットワーク (GAN) モデル⁽⁵⁾ (図 1(b)) とは異なるこのモデルは, アーキテクチャとしては非常にシンプルであるが, その画像変換 (生成) 能力は非常に高い。図 2 の (a) は, stable diffusion による画像生成例と, その時に入力した単語列である。たった

数単語しか用いてないが, 実写や人間が描いたかのような「それらしい」画像が生成されている。生成したい画像の特徴を上手く表現しているような単語を組み合わせれば, それに沿った出力をより高精度に行う事もできる。図 2(a) の右上の画像は, 質感や背景, 色, ポーズ, 光の加減などを 69 単語で指定し, その意図に沿った画像が生成された例である。

しかし画像によっては, 入力した単語の意図から外れたり, 「それらしくない」画像が生成されることもある。図 2(b) はその一例である。左上の腕を組んでいる女性は, よく見ると手の形状が崩れている (囲いの部分)。また, 右上のイラストは手首より先が 2 つ重なって描かれている (囲いの部分)。また, 図 2(b) 下 2 つの画像はシンプルな単円を生成しようとした結果であるが, 左の方は複数の円を重ね



(a) 拡散モデルのネットワーク構造概略図



(b) GANのネットワーク構造概略図

図1 拡散モデルとGANの概略図

て派手に描き、右の方は複数の歪んだ円になっている。

このように意図から外れる結果もあるが、総合して非常に強力な画像生成モデルである事は間違いない。ただ、何かを「付加する」外れ方が多く、「削除する」外れ方があまりない点が推察される。先の(b)の例でも、「単円」というシンプルな概念に対し、派手だったり歪んでいたりする画像を生成していた。ここから、拡散モデルにおいて、同じ画像変換（生成）タスクでありながら、ある概念について具

体化する事は「得意」であり、逆に抽象化する事は「不得意」である、という性質を持っているのではないかと考えられる。

なお、この性質は拡散モデルに限らず、GANによる生成においても確認できる。Yamane ら⁽⁶⁾はポリープの判別において深度画像を生成する際にGANモデルを用いているが、その結果FP（擬陽性）となった数多く、9割以上の再現率に比べ精度は8割弱と低くなっている。以上より、GANや拡散モデルのような画像生成モデルは、「ない」ものを「ある」とする（FPとなる）ような画像を生成しやすい性質、すなわち画像変換における具体化タスクと抽象化タスクに関する非対称な性質を持っていると考えられる。

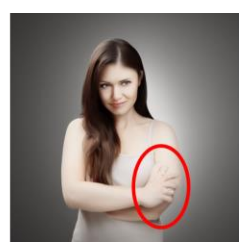
情報の抽象化・低次元化は、その情報の本質的な特徴を抽出することに等しい。画像に関する特徴抽出は、コンピュータによる画像の理解・認識においてより精度を高く保つべき部分である。そのため、先の性質についても、その詳細の解明が求められる。本研究では非対称な性質の解明を目指し、GANによる画像変換において革新的といわれ、現在の画像変換AIの基礎となっているpix2pix⁽⁷⁾を対象に、画像変換による具体化タスクと抽象化タスクにて見られる非対称的な性質について調査する。本稿ではまず、その性質の発現について実証する。



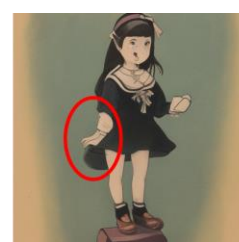
realistic, photo, statnding person



extremely detailed CG,...
(計69単語)



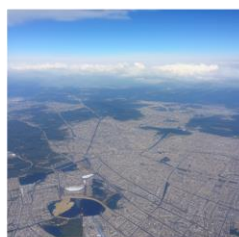
realistic, photo, statnding person



illust, standing girl



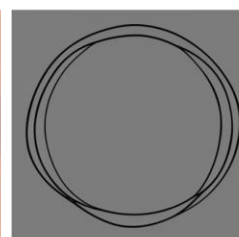
pictograph, simple



map, sapporo, simple, satellite



single circle



(a) 成功例（意図に沿う変換画像）

(b) 失敗例（意図から外れる変換画像）

図2 Stable Diffusionによる生成画像例（著者生成）

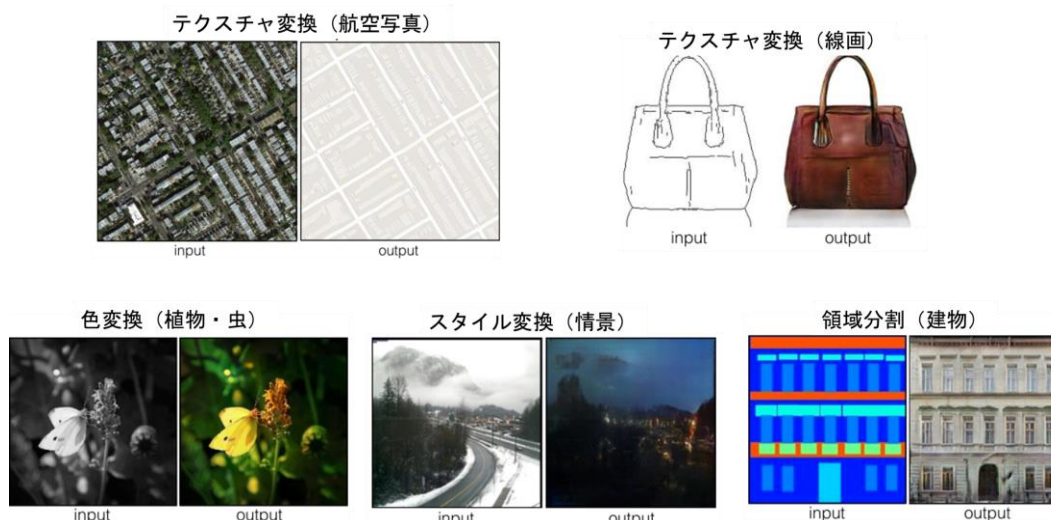


図3 pix2pix が対応可能なタスク例 (Isola ら⁽⁷⁾より転載, 一部改)

2. pix2pix とその性質

pix2pix は, 2017 年に Isola らによる論文で提案された GAN の一種であり, 入力データと識別時の比較データが同じものであるという制約付きの GAN (conditional GAN) である. 図 1(b)における「ノイズデータ」を「変換前画像」に変更して学習を行うモデルで, 生成器には Stable Diffusion などにも応用されている U-Net⁽⁸⁾, 判別機に PatchGAN を利用している. "Image-to-image translation"と名付けられたこのモデルは様々なタスクに対応しており, 図 3 のように領域分割, スタイル変換, 色変換, テクスチャ変換と言ったことが可能となっている. 各タスク, 左側の画像を入力とすると, 右側の画像が出力されるよう学習している. 出力画像はどれも入力画像をベースに変換が行われており, 近年の「生成」モデルのよりも, 元の画像の特徴を残した「変換」の要素が大きい.

pix2pix を改変・応用した研究は多くあり, 例えば Hollandi ら⁽⁹⁾は, 様々な種類の蛍光画像の領域分割のため, pix2pix によるスタイル変換を用いて元画像のマスク画像を生成する事で, Kaggle のコンペスコア (DBS) にて高い数値を出している. また Sato ら⁽¹⁰⁾も, 判別機の間層を用い新たな生成器を構築する Improved pix2pix を提案し, 細胞の蛍光画像から細胞膜と細胞核の領域分割を約 8 割の精度 (accuracy) で行っている. 他にも, 渡辺ら⁽¹¹⁾はデジタルイラストを自動でレイヤ分けし下塗りを行うシステムを構築しており, 山崎ら⁽¹²⁾は MIDI 譜面を画像化し, メロディから和音を生成する事を試みている.

このように, pix2pix は様々なタスクに様々な形で用いられ, 十分な性能を発揮している. しかし, 以上のような例においても, 先に述べたような性質が見られる. Yamane ら⁽⁶⁾の用いた GAN も pix2pix に依るものであるし, Sato らの例でも細胞核部分の生成精度の 7 割強と比べ, 細胞膜部分の生成精度に関しては 5 割強と低くなっており, 「面」の生成と「線」の生成において, より抽象度の高い後者の方が不得意であったと言える. 皆藤ら⁽¹³⁾による pix2pix を用いた画像の欠損部分の内挿実験においても, 欠損部分以外への補完すなわち「ない」ものを「ある」とした部分が見受けられる.

ところが, 元々がシンプルな画像の場合は, そういった性質が見受けられない. 植田ら⁽¹⁴⁾によるくずし字画像の裏抜け (紙に記述した文字の裏移り) 除去では, 複雑な画像においては元の文字も消えたりしているが, シンプルな文字においては綺麗に除去できている. 山本ら⁽¹⁵⁾のシンプルなレーダ画像を用いたコンクリートの内部欠損断面画像生成においても, 先のような性質は見受けられない. 以上より, この性質は, 元画像の複雑さと変換後のシンプルさに依存して発生すると考えられる.

本稿では, この性質が画像情報量 (エントロピー) に依存していると仮定し, 画像情報量と画像の具体化・抽象化 (テクスチャ変換) の精度との関係性を求めることで, 先の性質の実証を試みる.

3. 画像情報量に依る生成精度の非対称性の実証

本稿では, 複雑な実写の衛星写真と, シンプルに抽象化された地図画像との変換を行い, 衛星写真の

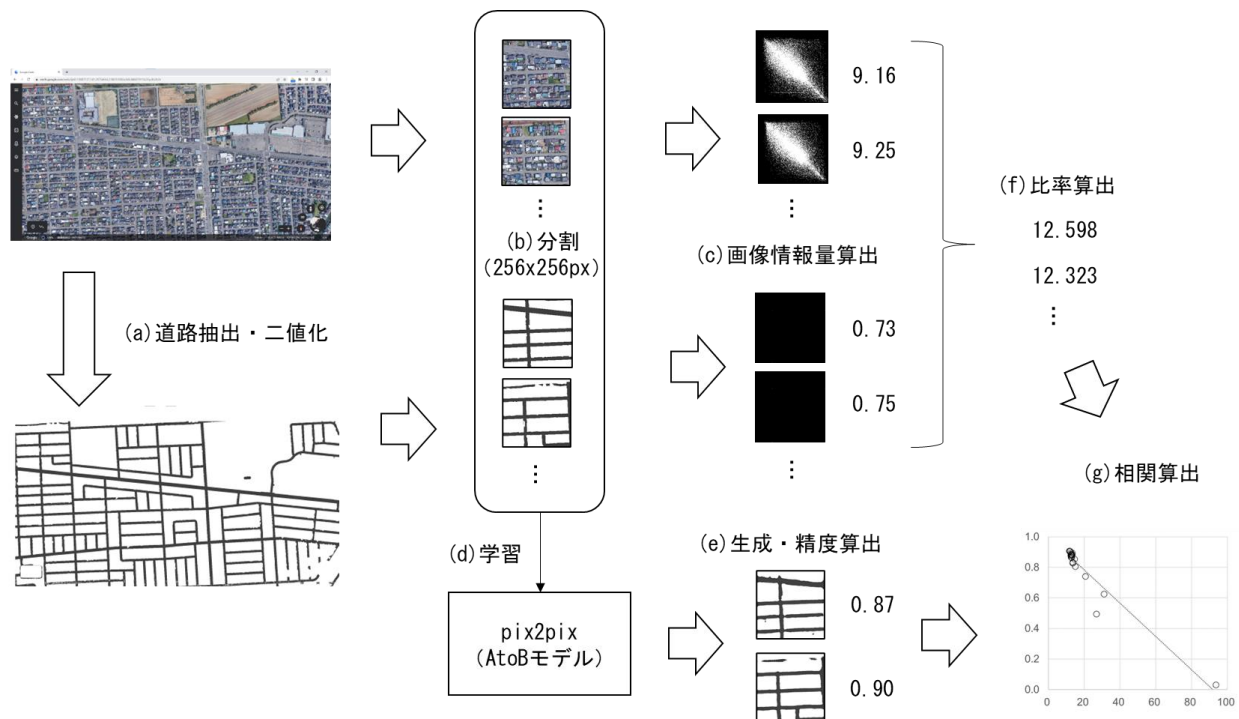


図4 実証実験手順 (AtoB モデル)

画像情報量と地図画像の画像情報量の比と、pix2pixによる地図画像の生成精度との相関を求めた。このような画像変換タスク（テクスチャ変換タスク）は、図3左上のようにpix2pixの原著でも触れられており、元の道路の特徴を残しつつ直線や曲線のような抽象的特徴を精確に取得し変換できるかどうかを検証する本研究に適していると考える。

まず、Google Earthを用い、検証に用いるエリアを選出した。ここでは、整備された住宅街、斜めに走る通り、曲線の道、空き地など様々な要素が含まれていたことから、手稲区星置近辺（緯度約43.14、経度約141.21近辺）とした（図4(a)上）。このエリアの衛星写真（約1792×768px）をGoogle Earthから取得し、地図画像を作成した。その際、衛星写真と、道路標示機能を用いた写真との差分を取り、2値化した（図4(a)下）。なお、ガウシアンや膨張などのノイズ除去フィルタは用いていない。

その後、これらをpix2pixへ入力するため256×256サイズに分割（図4(b)）した上で、各画像において画像情報量を算出した（図4(c)）。分割の結果、画像はそれぞれ横7枚×縦3枚の計21枚となった。

画像情報量はHaralick⁽¹⁶⁾らによる手法、すなわちGLCM（Gray Level Co-occurrence Matrix、同時生起行列）の各要素をピクセルの生起確率としたエントロピーとした。GLCMはグレースケール（輝度）画像におけるピクセル配置の類似性（同時生起）をその生起確率の行列で表したもので、様々な特徴量

に変換できることから、テクスチャ解析の他、リモートセンシングの場でも良く利用されている。本稿ではBT.601-7⁽¹⁷⁾によるグレースケール化を行った後、オフセット0（隣接したセルと）のGLCMを求めた。具体的には、画像の各ピクセル (x, y) におけるRGBベクトル $(r_{x,y}, g_{x,y}, b_{x,y})$ を用いて

$$G_{x,y} = \lfloor 0.298912 \cdot r_{x,y} + 0.586611 \cdot g_{x,y} + 0.114478 \cdot b_{x,y} \rfloor \quad (1)$$

にてグレースケール化した後に、基の輝度を行、隣接したピクセルの輝度を列とした輝度の出現頻度行列を作成し、各要素を総ピクセル数で割ることによりGLCMを得た。なお、生起確率が極端に小さくなる場合もあったため、 10^{-323} 以下は打ち切りで0とした。各画像におけるエントロピー E は、得られたGLCMの各要素を $P(i, j)$ とし、

$$E = \sum_i \sum_j P(i, j) \cdot \log P(i, j) \quad \dots\dots\dots (2)$$

で求めた。

また、(b)にて細分化した画像について、衛星写真を入力、地図画像を出力としたpix2pixモデル（AtoBモデル、抽象化タスクに対応）と地図画像を入力、衛星写真を出力としたpix2pixモデル（BtoAモデル、具体化タスクに対応）の学習を行った（図4(d)）。学習エポック数はどちらも100、学習のバッチサイズは1とした。

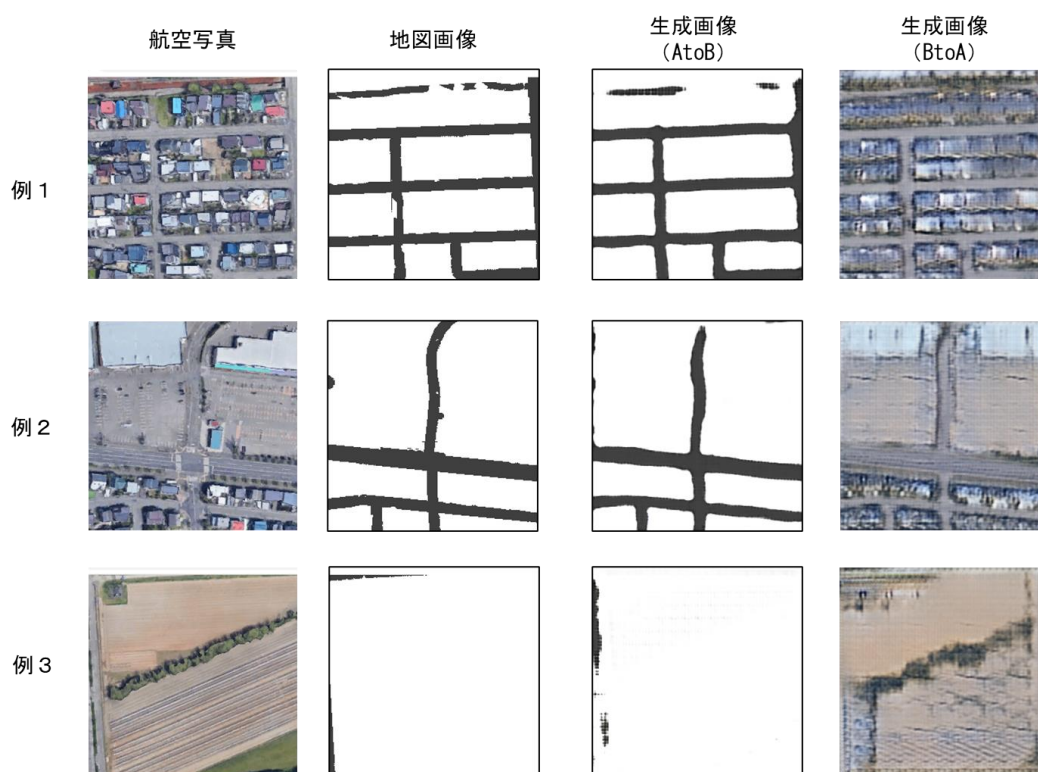


図 5 生成画像例

学習後、同じ画像を再入力して画像を生成させ、正解画像のピクセル (x,y) における RGB ベクトル $C_{t,x,y}$ と生成画像における同 RGB ベクトル $C_{g,x,y}$ とのコサイン距離の平均値

$$sim = \frac{1}{x \cdot y} \sum_x \sum_y \frac{C_{t,x,y} \cdot C_{g,x,y}}{|C_{t,x,y}| \cdot |C_{g,x,y}|} \quad \dots\dots (3)$$

で生成精度を求めた。また更に AtoB モデルにおいては、より分かりやすい指標として、正解の地図画像と生成した地図画像との各ピクセルにおける道路描画の有無（ピクセルの色が白（255, 255, 255）なら無、それ以外なら有）を求め、正解画像での有無を True/False, 生成画像での有無を Positive/Negative とした時の適合率（precision）・再現率（recall）・F 値を求めた。

最後に、AtoB モデルでは衛星写真と地図画像の画像情報量の比（衛星写真÷地図画像）と F 値、BtoA モデルでは同比と類似度との相関係数を算出し、それぞれどのような関係があるか確かめた。

4. 実験結果

まず、各モデルにおける画像生成結果例を図 5 に示す。例 1 は地図らしい縦横に走る直線的な道路の部分で、例 2 は曲線や斜めに走る道路の部分である。

どちらの例でも、「それらしい」画像が生成できていることが分かる。ただ、AtoB モデル（抽象化タスク）による地図画像をみると、全体的に線が太く歪み、欠損部分を補完しているのも見て取れる。例 3 は情報量比が最大の画像（No. 13）で、ほぼ道路は存在しないが、AtoB・BtoA とともに「それらしい」ものは生成されているのが分かることから、学習面に関してはある程度妥当に行われたことが分かる。

次に、各画像情報量とその比、AtoB モデルにおける適合率・再現率・F 値、BtoA モデル（具体化タスク）における類似度の値を表 1 に示す。表より、AtoB モデルにおいて再現率に比べ適合率が低い。これは、本来道路でないピクセルに道路のピクセルを生成することが多かったと言え、やはり「ない」ものを「ある」とする、FP（擬陽性）のような外し方が多い事が分かる。

BtoA モデルにおいては、類似度が非常に小さくなっている。これは、本来具体化タスクが正解と完全に一致する画像を生成するタスクではなく、あくまで「それらしい」画像を生成するタスクであることに因ると考える。図 5 右のように生成した画像に「それらしさ」がある点（地図の白い面に家らしきものが描かれている等）、また今回は画像情報量比との相関を求めるため相対的な差が分かればよい事が

表 1 各画像における諸値

ID	衛星写真 情報量	地図画像 情報量	情報量比	AtoBモデル			BtoAモデル
				適合率 (precision)	再現率 (recall)	F値	類似度
1	9.161	0.727	12.598	0.793	0.956	0.867	2.925.E-03
2	9.253	0.747	12.393	0.830	0.976	0.897	3.162.E-03
3	9.352	0.751	12.446	0.809	0.981	0.887	3.293.E-03
4	8.403	0.314	26.751	0.343	0.890	0.495	2.163.E-03
5	9.030	0.769	11.744	0.853	0.969	0.907	3.099.E-03
6	9.265	0.795	11.655	0.846	0.979	0.908	3.167.E-03
7	8.938	0.748	11.942	0.838	0.984	0.905	2.942.E-03
8	9.213	0.724	12.723	0.800	0.986	0.883	3.175.E-03
9	9.276	0.702	13.215	0.810	1.000	0.895	3.245.E-03
10	8.835	0.651	13.574	0.807	0.862	0.834	2.687.E-03
11	9.107	0.736	12.376	0.819	0.947	0.878	3.053.E-03
12	9.305	0.701	13.271	0.799	0.990	0.884	3.307.E-03
13	8.311	0.089	93.638	0.018	0.236	0.033	2.021.E-03
14	8.870	0.596	14.890	0.722	0.913	0.806	2.492.E-03
15	9.286	0.718	12.924	0.792	0.978	0.875	3.327.E-03
16	8.724	0.281	31.050	0.459	0.981	0.625	2.443.E-03
17	8.477	0.586	14.460	0.765	0.968	0.854	1.000.E-05
18	9.313	0.715	13.023	0.796	0.996	0.885	1.000.E-05
19	8.994	0.664	13.552	0.739	0.940	0.828	3.055.E-03
20	8.650	0.420	20.604	0.594	0.983	0.741	2.298.E-03
21	9.091	0.703	12.927	0.777	0.974	0.865	3.715.E-03

ら、不備等とせずそのまま扱う。

最後に、AtoB モデルにおける画像情報量比と F 値、BtoA モデルにおける画像情報量比と類似度についての相関係数を、図 6 に示す。横軸が画像情報量比、縦軸が F 値もしくは類似度である。

AtoB モデルにおいては、相関係数が約-0.96 となり、非常に強い負の相関があった。このことから、画像情報量比が上がる（より複雑な衛星写真をよりシンプルな地図画像にする）ほど、生成された画像は正解の地図画像から離れていくことが分かった。その逆の BtoA モデルでは相関係数が-0.30 と弱い負の相関であり、AtoB モデルのような傾向が薄弱であることが分かった。

以上より、pix2pix による画像変換において、具体化タスク・抽象化タスクにおける画像生成精度に非対称な性質があることが実証できた。

5. 考察

この性質について、Stable Diffusion や pix2pix といった画像生成モデルにおける生成器として頻

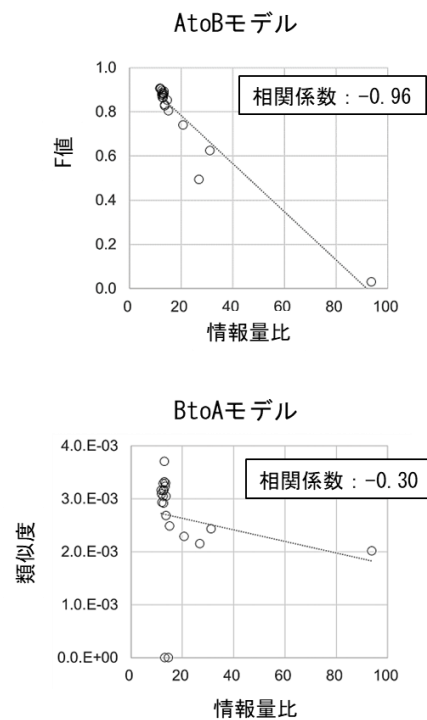


図 6 各モデルでの相関算出結果

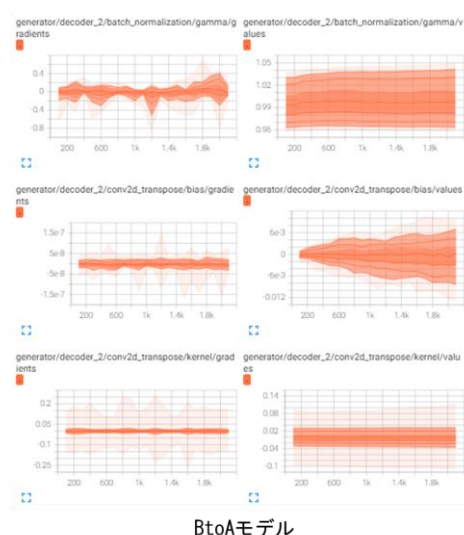


図7 バッチ正規化係数 γ (上), バイアス (中), カーネル (下) の値と勾配の時間変化 (U-Net デコーダ部分 2 層目, TensorBoard により表示)

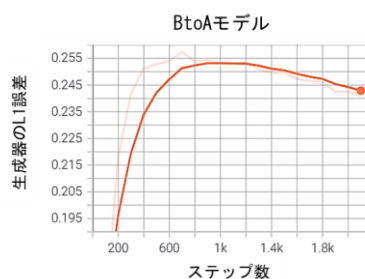
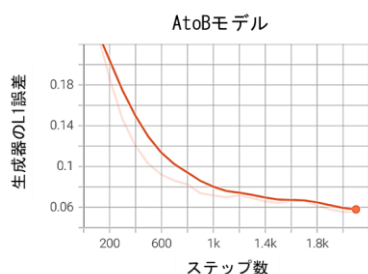


図8 U-Net 学習時のステップ毎の L1 誤差 (TensorBoard により表示)

繁に用いられている U-Net⁽⁷⁾に焦点を当てて考えてみる。U-Net は encoder-decoder モデルの一種で、エンコーダと同じ層数を持つデコーダを用意し、ある階層におけるエンコーダの出力を同じ階層のデコーダにも入力する (skip-connection と呼ばれている) ことで、入力画像のディテールを保持しつつデコードが行えるモデルである。

U-Net の各層におけるネットワークの重みに関するパラメータ (カーネル・バイアス・バッチ正規化係数) の値と勾配の時間変化について AtoB モデルと BtoA モデルとを比較したところ、明確な差異はみられなかった。例として、図 7 にデコーダ 2 層目の様子を一部示すが、微細な差異しか見受けられない。残りの層 (全 8 層) においても同様であった。

しかし、最適化における正則項である L1 誤差の減少スピードにおいては、両モデルで非常に大きな差が出た。図 8 は左が AtoB モデル、右が BtoA モデ

ルにおけるステップ毎の L1 誤差を示している。AtoB モデルでは急速に減少・収束していくのに対し、BtoA モデルでは上昇後ゆるやかに減少している。このことから、AtoB モデルは BtoA モデルに比べ非常にスパース (疎) な重みを持ちやすいといえる。

だが、L1 誤差の本質であるスパース性を持った解の導出は、重みの解空間における次元の削減を行うものであり、つまり「細かい点を無視してでも全体を整えている」ものであるとも考えられる。それは過学習している訳ではなく正則化の目的が果たされている訳だが、こと U-Net における抽象化タスクに限ると、skip-connection による元の画像情報が入力として存在するため、「ない」より「ある」方がまだ良いという方向すなわち情報量が大きくなる方に収束し易くなり、その結果、情報量の差が大きいほど生成結果に先の性質が顕著に表れてしまうのではないかと推察される。今後、U-Net 自体の

持つ性質を解明できれば、画像の抽象化・低次元化の精度向上に活用できると考える。

6. おわりに

本研究では、画像変換 AI における具体化タスクと抽象化タスクとの非対称な性質について、現在の画像変換 AI の基礎となっている pix2pix によるテクスチャ変換を対象に実証を行った。

衛星写真を地図画像へ変換する AtoB モデル（抽象化タスク）とその逆の BtoA モデル（具体化タスク）について、各画像の GLCM を求めた後に情報量を算出し、衛星写真の値と地図画像の値との比と、pix2pix による生成画像の精度との相関を求めたところ、抽象化タスクを担う AtoB モデルにおいては画像情報量比が上がるほど精度が下がる負の相関があり、具体化タスクを担う BtoA モデルでは弱い負の相関しかないことが分かった。すなわち、pix2pix は具体化タスクより抽象化タスクの間に非対称な性質がある事が実証できた。

この性質について、画像生成モデルに頻繁に用いられる U-Net の L1 誤差に着目したところ、AtoB モデルにおける収束スピードが BtoA モデルに比べ早い事が分かった。ここから、抽象化タスクでは重みの解空間がスパースになり、U-Net の構造が持つ skip-connection の影響もあって情報量が大きくなる方に学習が進んでいると推察される。今後、そのような U-Net の複雑な性質を解明し、画像情報の抽象化・低次元化の精度向上に活用したい。

参考文献

- (1) Midjourney: Midjourney, 2023/1/26, <https://midjourney.com>.
- (2) R. Rombach, A. Blattmann et al.: High-Resolution Image Synthesis with Latent Diffusion Models, arXiv:2112.10752, 2022.
- (3) Anlatan: NovelAI - the GPT-powered AI storyteller, 2023/1/26, <https://novelai.net>.
- (4) J. S-Dickstein, E. A. Weiss et al.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics, arXiv:1503.03585, 2015.
- (5) I. J. Goodfellow, J. P-Abadie et al.: Generative Adversarial Networks, arXiv:1406.2661, 2014.
- (6) H. Yamane, S. Fukui et al.: Automatic Generation of Polyp Image using Depth Map for Endoscope Dataset, Procedia Computer Science, Vol.192, pp.2355-2364, 2021.
- (7) P. Isola, J-Y. Zhu et al.: Image-to-Image Translation with Conditional Adversarial Networks, arXiv:1611.07004, 2018.
- (8) O. Ronneberger, P. Fischer and T. Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv:1505.04597, 2015.
- (9) R. Hollandi, A. Szkalitsy et al.: nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer, Cell Systems, Vol.10, pp.53-458, 2020.
- (10) M. Sato, K. Hotta et al.: Segmentation of Cell Membrane and Nucleus by Improving Pix2pix, Proceedings of the 11th International Joint onference on Biomedical Engineering Systems and Technologies, Vol.4, pp.216-220, 2018.
- (11) 渡邊 優, 阿倍 博信: Smart Layer Splitter: pix2pix を用いた デジタルイラスト制作の色塗り工程における 自動レイヤ分けシステム, 情報処理学会論文誌 デジタルコンテンツ, Vol.9, No.1, pp.21-33, 2021.
- (12) 山崎 健一, 坂 知樹, 鎌田 洋: 深層学習を用いた主旋律に基づく和音生成, 映像情報メディア学会誌, Vol.77, No.1, pp.135-140, 2023
- (13) 皆藤 優太, 田村 仁: 深層学習を用いた輪郭線情報に着目した画像修復, 情報処理学会第 81 回全国大会講演論文集, pp.185-186, 2019.
- (14) 植田ちひろ, 藤岡寛之, 日高章理: Pix2Pix を用いた古典籍くずし字画像の裏抜け除去, 情報処理学会第 81 回全国大会講演論文集, pp.187-188, 2019.
- (15) 山本 佳士, 光谷 和剛ほか: 準 3 次元情報を用いた pix2pix によるレーダ 画像からの内部欠陥の幾何情報推定, AI・データサイエンス論文集, Vol.3, No.2, pp.1042-1052, 2022.
- (16) R. M. Haralick, K. Shanmugam et al.: Textural Features for Image Classification, IEEE Transactions on Systems, Man, and Cybernetics, Vol.3, No.6, pp.610-621, 1973.
- (17) Mathworks: MATLAB - img2gray, 2023/1/26, <https://jp.mathworks.com/help/matlab/ref/img2gray.html>.