輪郭線と色からなるカラー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,masaki.aono.ss}@tut.jp

あらまし　本研究では, pix2pixHD をベースとした手法を用いて, 輪郭線と色からなるカラー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. 近年の深層学習を用いた画像生成の進歩はセグメンテーションマップやテキストからの高品質な画像生成を可能にしたが, これらの手法ではラベルや言語に依存するために汎化性能やユーザビリティが低く, 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や言語に依存しない入力形式として輪郭線と色からなるカラースケッチ画像を用いてフォトリアルな画像を生成する手法を提案する. さらに生成モデルとして Diffusion Model に基づく U-Net と DPT-Large を導入した手法を提案し, FID 評価尺度と KID 評価尺度, アンケートによる評価にて精度改善したことを報告する.

キーワード　深層学習, GAN, 画像生成

1　はじめに

近年, 深層学習による画像生成の活発化により非常に高品質な画像が生成できるようになりつつある. 条件を設けない単純な生成だけでなく, 条件を与えた画像生成も非常に高品質に行えるようになっている. 条件をユーザが与えることで欲しい画像を生成しやすくなるため, 画像素材の作成や画像制作の補助ツールとしての応用が可能となる. また, 高度なスキルを用いることなく画像生成が可能となるほか, 従来の画像制作ソフトと比較して短時間での画像制作の実現が考えられる. 画像生成時に設ける条件としては, セグメンテーションマップやテキストを条件とした画像生成に関する研究が活発に行われている. 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, ラベルとして事前に設定して学習させた物しか生成することができないという問題がある. さらに, 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, 猫の画像を生成したいときに「動物」というラベルしか存在しない場合, 猫の画像が生成されるとは限らない. しかし, 数多存在するすべての動物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. セグメンテーションマップからの生成での問題を解決するために, 近年ではテキストからの画像生成が活発になっており, 簡単なテキストから画像生成が可能である上により多様な画像を生成可能になっている. 一方で, 言語に依存することとユーザが想像している画像が生成されるとは限らず何度も生成をやり直す必要がある場合があるという短所も存在する. これらの問題を解消するために本研究では, ラベルや言語に依存しない方法として, 図 1 に示すような, 物体の輪郭線と色からなるカラースケッチ画像を入力形式として用いた手法での画像生成の精度向上を目的とする.

(a) 線画　**(b)** 色画像　**(c)** カラースケッチ　**(d)** 生成画像

図 1: カラースケッチからの画像生成. (a) は物体の輪郭線を描いた線画, (b) は (a) に対応する塗り絵を行った色画像. (c) は (a) の下のレイヤーに (b) を置き, 合成することにより得られるカラースケッチ画像. 画像生成モデルに (c) を入力することで (d) のフォトリアルな生成画像が得られる.

カラー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 [1] 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) [2] を利用した画像変換手法の一種である. GAN は Generator と Discriminator の 2 つのネットワークから構成される. Generator は目的とする生成物を生成するためのネットワークで, Discriminator は Generator が生成した生成物と本物を識別するためのネットワークとなっている. Generator は Discriminator を騙すように, Discriminator は Generator の生成物を見抜けるように学習を行い, 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 pix2pixHD [3] をベースとして Diffusion Model [4] で用いられる U-Net を Generator として導入し, DPT-Large [5] を用いた深度推定結果を活用する手法を提案し, FID 評価尺度と KID 評価尺度, アンケートによる評価にて精度改善したことを示す.

2 関連研究

画像生成は, GAN [2] 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. 近年では Diffusion Model [4] を用いた手法によりさらに高品質な生成が可能になっている. また, スケッチやペインティングを入力とした画像生成に関する研究も提案されている. 以下では, GAN と Diffusion Model を用いた画像生成及びスケッチやペインティングからの画像生成に関する研究について述べる.

2.1 GAN を使用した画像生成

通常の GAN はノイズのみを入力とするが, GAN から派生した手法として条件を入力とする Conditional GAN [6] が存在する. これによって, 条件に対応した生成ができる. Isola らの pix2pix [1] 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解し, 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 に対しても 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, 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 [3] 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 Discriminater を用意して小さなスケールに変換した画像に対しても Discriminater を適応できるように工夫されている. これにより pix2pixHD は, 2048x1024 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.

2.2 Diffusion Model を使用した画像生成

画像生成は, GAN をベースとする手法が長らく主流となっていたが, Ho らは, 確率拡散モデルを用いた高品質な画像生成手法である DDPM(Denoising Diffusion Probabilistic Models) [4] を提案した. この手法は, 画像に対して小さなノイズを少しずつ足していくことでノイズ画像に変換する過程を考え, その逆変換によってノイズ画像から画像を生成する方法である. より具体的には, ノイズ画像に対してかかっているノイズをニューラルネットワークで予測し, それを繰り返すことで少しずつノイズを取り除くことで画像を生成を行う. Diffusion Model は学習を二乗誤差最小化で定式化でき, GAN のように Discriminator を用いたりしないことから GAN と比較すると複雑ではなく, GAN で発生しやすい勾配消失やモード崩壊などの問題の影響を受けないという利点がある. また, GAN と比較して多様なデータの生成に強いことも特徴として挙げられる [7]. しかし繰り返しノイズを予測することから GAN と比較して生成時間がかかる [8]. さらに, 高解像度の画像を生成する場合には GAN と比較してより巨大なネットワークが必要となるため, モデルパラメータ数と計算量が増大する問題がある [7].

Diffusion Model では, 手法そのものの工夫以外にもネットワークアーキテクチャの工夫が行われている. 多くの Diffusion Model は DDPM [4] で有効性が示された U-Net アーキテクチャをベースとして改良されている. U-Net [9] は, 医療用画像のセマンティックセグメンテーション向けに提案されたスキップ接続を特徴とする, Encoder-Decoder ネットワークである. DDPM ではこれに対して, 工夫を施した U-Net を提案している. 具体的には, Diffusion Model として用いるために時点情報を表す埋め込み表現の導入, Wide ResNet [10] の導入, Attention の導入, Group Normalization [11] の導入がされている. これらの工夫により Diffusion Model での画像生成が実現された. これに対して Dhariwal らは, この U-Net アーキテクチャをベースとしてアーキテクチャ探索を行い, 高品質かつ高解像度の画像生成を実現した ADM(Ablated Diffusion Model) を提案した [7]. 具体的にはモデルの深さ, Attention 機構のヘッド数, ResNet Block に対して対照実験が行われ, モデルの探索が行われた. この研究により, Diffusion Model が当時の最先端であった BigGAN [12] の精度を上回ることが示された.

2.3 スケッチやペインティングからの画像生成

近年, スケッチやペインティングを入力として画像生成を実現する手法も提案されている. Isola らの pix2pix [1] では, Edges to Photo として輪郭線からのフォトリアルな画像生成を提案した. Edges to Photo は HED (Holistically-Nested Edge Detection) [13] を用いて生成した輪郭線画像からフォトリアルな画像に復元するタスクで靴や鞄を輪郭線画像から生成する試みが行われた. Wang らは PITI(pretraining-based image-to-image translation) [14] を用いた画像変換タスクの一つとして Sketch-to-image synthesis を提案した. Sketch-to-image synthesis についても HED [13] を用いて抽出したスケッチ画像をフォトリアルに変換するというものである. Meng らはペインティングからの画像生成手法として SDEdit(Stochastic Differential Editing) [15] を提案した. この手法では, 画像に対してメディアンフィルタを適応後, 減色することで画像と対応するペインティング画像を生成し, この画像を入力として Diffusion Model によるペインティングからフォトリアルな画像への変換を評価している. Zhang らは, 学習済みの Diffusion Model の一つである Stable Diffusion [16] に対して条件を追加して制御するネットワーク構造である, ControlNet [17] を提案した. この手法により, Stable Diffusion に対して様々な画像条件付きの生成が可能となった.

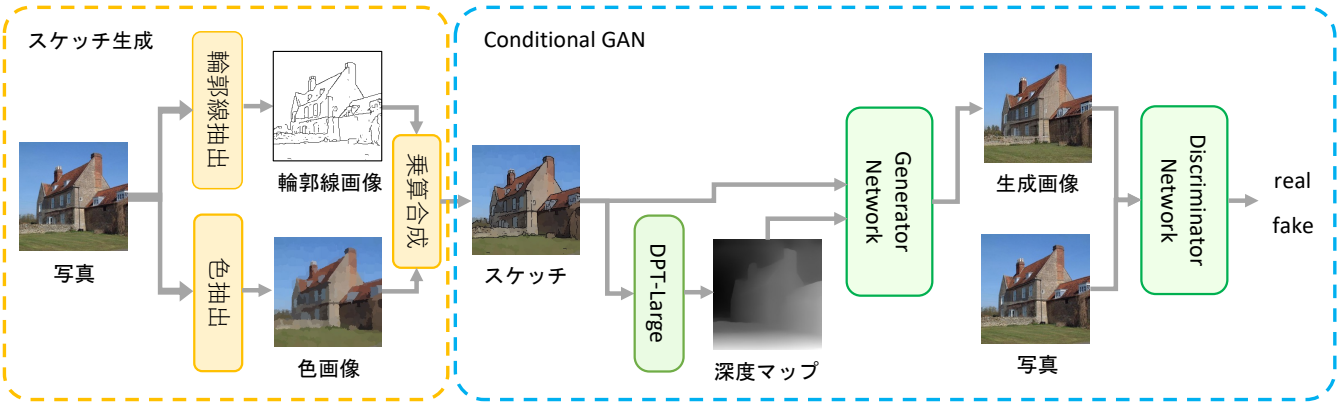

図 2: 提案手法の流れ. スケッチ生成部により, 写真に対応する人が描いたようなスケッチを自動生成する.Conditional GAN 部ではスケッチから写真の状態に復元することを目的として学習を行う.

3 提案手法

提案手法の流れを図 2 に示す. ユーザは「生成したい画像の輪郭線を描いた線画」と「線画に対応する色画像」を描き, 線画の下レイヤーに色画像を配置して合成したものを入力とする. この線画と色画像は人が描くことを想定しているが, 学習時に写真と対応する人が描いたスケッチを大量に用意することは困難である. そこで本研究では線画と色画像を写真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

本研究では, スケッチの生成方法としてスケッチ 1 とスケッチ 2 の 2 つの手法を用いる.

スケッチ 1 では, 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 [18] を適応し細線化を行う. 色の抽出方法は, スーパーピクセルのアルゴリズムの SEEDS (Superpixels Extracted via Energy-Driven Sampling) [19] 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色画像とする.

スケッチ 2 では, 線の抽出方法として HED [13] を用いる. pytorch-hed [20] を用いて輪郭線を抽出し, Zhang-Suen アルゴリズム [18] を適応し細線化を行うことで, 輪郭線を抽出する. 色の抽出方法には, SDEdit [15] の Stroke Painting と同様の手法を用いる. 具体的には, カーネルサイズ 23 のメディアンフィルタをかけた後に Pillow [21] の quantize を用いて, 6 色に減色したものを色画像とする.

3.2 Conditional GAN

本研究では, カラースケッチとカラースケッチから学習済みモデルを用いて得られた深度マップを入力として pix2pixHD の Generator 部に変更を加えたモデルを提案する. 具体的には Generator を Diffusion Model に基づく U-Net に変更し, 深度マップは DPT-Large を用いて推定する. 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 Diffusion Model に基づく U-Net と DPT について説明する.

3.2.1 Diffusion Model に基づく U-Net

2.2 節で述べたように多くの Diffusion Model では DDPM [4] で有効性が示された U-Net がベースとして使用されている. 本研究では, アーキテクチャ探索により高精度化が行われた ADM [7] で使用されている U-Net を用いる. しかし, GAN では Diffusion Model のための時点情報を扱う機能を必要としないことと, モデルパラメータ数が約 5.5 億 (554M) と非常に巨大であることから学習に膨大な時間を必要とし, そのまま使用することは出来ない. そこで時点情報を扱う機能の削除とモデルパラメータ数の削減を行い, pix2pixHD の Generator として導入する. 具体的には, 5 回のダウンサンプリング数を 3 回まで減らし最小の特徴量マップのサイズを 1/32*1/32 から 1/8*1/8 のサイズまでのダウンサンプリングまでに削減, 特徴量チャンネル数をボトルネック部を除き削減, ボトルネック部と 1/8*1/8, 1/16*1/16, 1/32*1/32 の特徴量マップの領域で ResNet Block の後に配置されていた Attention Block をボトルネック部のみに削減, 時点情報の埋め込み表現を計算する機構と ResNet Block から時点情報を扱う機構の削除を行った. これらの操作によってパラメータ数を約 5000 万 (50.5M) まで削減した. 図 3 にパラメータ削減と時点情報機能の削除を行った Generator のネットワーク図を示す.

3.2.2 DPT(Dense Prediction Transformers)

DPT [5] とは Ranftl らが提案した, Vision Transformer(ViT) [22] を使用したセグメンテーションモデルである. 画像のセマンティックセグメンテーションと単眼深度推定において, CNN をベースとするモデルと比較して高精度な推定を実現させている. 特に単眼深度推定において, DPT は物体の境界でより詳細な出力を生成する傾向があることが示されている [5]. 本研究では, ViT-Large をベースとする DPT-Large の学習済みモデルを用いてスケッチから深度推定を行う. そして得られた深度マップをカラースケッチとともに GAN の入力とする.

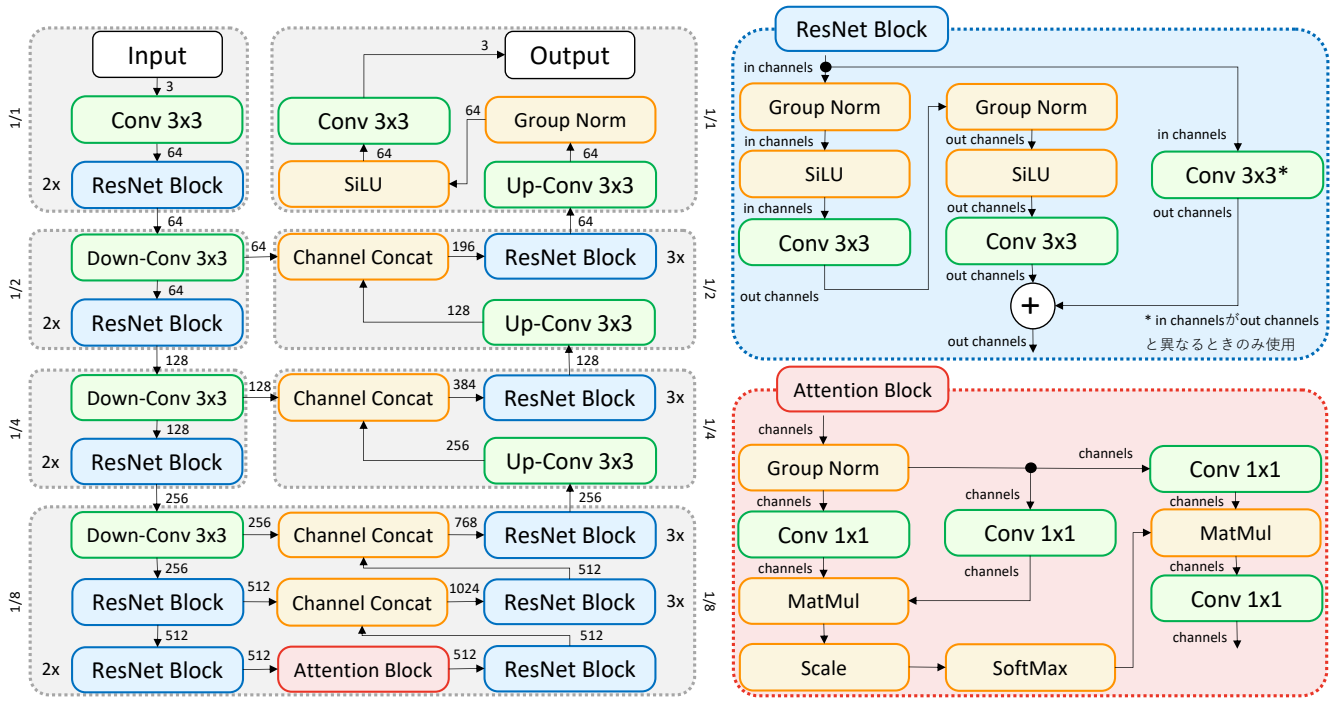

図 3: Diffusion Model に基づく U-Net (提案手法 2 および提案手法 3 の Generator)

3.2.3 ベースライン手法

本研究ではベースライン手法として, pix2pix [1] を用いる手法と pix2pixHD [3] を用いる手法の 2 種類を使用する. これらのベースラインモデルを前者はベースライン手法 1, 後者はベースライン手法 2 と呼称する. ベースライン手法 2 では, pix2pixHD の Global generator network のみを Generator として使用する.

3.2.4 提案手法

本研究では提案手法として, ベースライン手法 2 をベースとして変更を行った 3 つのモデルを提案する. 提案手法 1 は, pix2pixHD に対して DPT-Large により推定した深度マップをカラースケッチに加えて追加の入力とするモデルであり, pix2pixHD+DPT-Large と呼称する. 提案手法 2 は, 図 3 に示した Diffusion Model に基づく U-Net を Generator とした pix2pixHD であり, pix2pixHD-Unet と呼称する. 提案手法 3 は提案手法 1 と提案手法 2 を組み合わせ, 提案手法 2 に対して DPT-Large により推定した深度マップを追加の入力としたモデルであり, pix2pixHD-Unet+DPT-Large と呼称する.

4 実　　験

4.1 データセット

本研究では, ADE20K データセット [23] と LHQ データセット [24] 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デットである. 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 LHQ データセット

LHQ データセットは自然風景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 90,000 枚から構成される. 画像サイズは 256 × 256 である. 本研究では, 学習用 81,000 枚, 検証用 9,000 枚に分割し, 実験を行った.

4.2 評 価 指 標

評価指標には Fréchet Inception Distance(FID) [25] と Kernel Inception Distance(KID) [26] 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. FID スコアと KID スコアの算出には clean-fid [27] を用いた. また, 各モデルのパラメータ数についても評価する. パラメータ数は Generator のパラメータ数を算出し, DPT-Large 使用時はそのパラメータ数も加えて評価する.

4.3 アンケートによる評価

本研究では, 特徴距離の評価に加えてアンケートによる人による評価も実施した. アンケートでは, 10 枚の手描きのカラースケッチに対して各モデルを用いて生成を行い, 「写実的だと思う順番に順位づけをしてください。」と説明し, 各手法の生成結果を順位付けする方法を用いた. さらに, アンケート結果から各手法の票が各順位までに含まれている割合を算出した.

4.4 モデルパラメータ

バッチサイズは 8, epoch 数は ADE20K データセット学習時は 50, LHQ データセット学習時は 15 とした. また, 学習率は 0.0001 としている. 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 と同様の値で実験を行った.

表 1: 各モデルの実験結果と各モデルのパラメータ数. FID スコアと KID スコアは生成結果と Groud Truth 間の特徴量距離を表す. パラメータ数については Generator のパラメータ数である. ただし, DPT-Large を使用する手法についてそのパラメータ数を加えて表示している.

			スケッチ 1 256 x 256				スケッチ 2 256 x 256			
			ADE20K		LHQ		ADE20K		LHQ	
Method		Params(M)	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓
pix2pix	(ベースライン手法 1)	54.40	48.18	0.0202	31.67	0.0148	116.79	0.0715	62.05	0.0396
pix2pixHD	(ベースライン手法 2)	45.59	27.74	0.0055	23.44	0.0110	53.56	0.0182	30.87	0.0140
pix2pixHD + DPT-Large	(提案手法 1)	389.05	33.01	0.0077	26.34	0.0132	57.93	0.0212	36.11	0.0176
pix2pixHD-Unet	(提案手法 2)	55.00	17.65	0.0013	8.56	0.0026	29.96	0.0047	9.77	0.0020
pix2pixHD-Unet + DPT-Large	(提案手法 3)	398.46	**16.66**	**0.0009**	**6.56**	**0.0018**	**29.35**	**0.0046**	**8.89**	**0.0018**

表 2: アンケートにおける各種法の票数

		票数					Ratio@N(N 位以上の表の割合)			
Method		1 位	2 位	3 位	4 位	5 位	Ratio@1	Ratio@2	Ratio@3	Ratio@4
pix2pix	(ベースライン手法 1)	1	1	16	12	70	0.01	0.02	0.18	0.30
pix2pixHD	(ベースライン手法 2)	11	16	33	32	8	0.11	0.27	0.60	0.92
pix2pixHD + DPT-Large	(提案手法 1)	16	14	28	37	5	0.16	0.30	0.58	0.95
pix2pixHD-Unet	(提案手法 2)	36	34	6	11	13	**0.36**	0.70	0.76	0.87
pix2pixHD-Unet + DPT-Large	(提案手法 3)	36	35	12	15	2	**0.36**	**0.71**	**0.83**	**0.98**

4.5 実験結果

表 1 は ADE20K データセットと LHQ データセットでの実験結果と各モデルのパラメータ数である. 表 1 より, pix2pixHD の Generator を Diffusion Model に基づく U-Net した提案手法 2 は pix2pixHD と比較して両方のデータセットの各スケッチ生成方法で精度が向上していることがわかる. さらに DPT-Large を用い深度マップを追加の入力とした提案手法 3 はそれぞれのデータセットで最も高い精度となった. しかし, pix2pixHD に対して深度マップを追加の入力として与えた提案手法 1 に関しては pix2pixHD に対して精度が低下した.

図 4 は LHQ データセットのスケッチ 1 での生成例である. 図 4 より pix2pix では細かな質感の再現が上手くできておらず, ぼやけた印象となっていることが確認できる. pix2pixHD は, 質感については pix2pix よりも良いものの輪郭がはっきりしない画像となっていることが確認できる. 提案手法についてはこれらの問題が軽減されており, 質感や輪郭のぼやけが改善されていることが確認できる. しかし, 建物が含まれる例については提案手法 2 と提案手法 3 においても余計な質感が付与されていたりする例が確認できた. 図 5 の LHQ データセットのスケッチ 2 での生成例については, pix2pix では図 4 での生成例と同様に質感がのっぺりとしていることが確認でき, pix2pixHD についても輪郭のぼやけが確認できる. これに対して提案手法では物体の輪郭がはっきりとし, 細かな質感が生成されていることが確認できる. さらに, Ground truth と比較しても遜色ない品質で復元できていることが確認できる. しかし, 人工物が含まれる例については上手く生成できない箇所が含まれる例も確認できた.

また, 図 6 はペイントソフトを用いて人が描いたスケッチからの生成結果であり, LHQ データセットでスケッチ 2 での学習を行った各モデルを用いて生成を行った. 図 6 より, 人が描いたスケッチからの生成においてもフォトリアルな画像を生成できていることが確認できる. 特に, 提案手法 2 と提案手法 3 は草木や岩肌の質感や水面の表現が上手く生成できていることが確認できる. しかし, 図 4 の一番下の例のように, 空部分に不自然なテクスチャが現れるケースも確認できた.

4.5.1 アンケート結果

図 6 の結果を用いて 10 名に対してアンケートを実施した結果が, 表 2 である. Ratio@N は全体の票に対して N 位以上に選ばれた割合を表す. 表 2 より提案手法 3 は最も上位に選ばれる票が多く, 下位に票が入ることが少ないことが分かる. ただし, 1 位に選ばれた数は提案手法 2 と提案手法 3 で同数の 36 票であることが確認できる.

4.6 考　　察

実験結果より, Diffusion Model に基づく U-Net を導入した提案手法 2 は pix2pix 程度のパラメータ数を保ちながら高精度化することに成功した. これにより, Diffusion Model に基づく U-Net が Diffusion Model だけでなく GAN の Generator としても効果的であることが確認できた. これは, Wide ResNet や Attention, Group Normalization が GAN に対しても有効であったことが考えられる. また, pix2pixHD の Generator がダウンサンプリング後のボトルネック部に ResNet が配置されているのに対し, Diffusion Model に基づく U-Net ではダウンサンプリング毎に ResNet が配置されており, 大きい解像度の特徴量マップに対しても ResNet を適応させていることが精度の向上につながっていると考えられる. さらに, DPT-Large を用いて深度マップを追加で与える拡張を行った提案手法 3 は, 提案手法 2 よりさらに精度を改善した. これは, 深度マップにより同じ物体の領域が近い深度でグルーピングされたことが, 質感の描き分けや空間表現の向上につながったと考えられる. しかし, pix2pixHD に対して深度マップを追加の入力として与えた提案手法 1 については, 精度が低下する結果となった. これは,

図 4: LHQ データセットでのスケッチ 1 の各手法の生成例. Ground truth は実際の写真, Color Sketch は入力するカラースケッチである.

図 5: LHQ データセットでのスケッチ 2 の各手法の生成例. Ground truth は実際の写真, Color Sketch は入力するカラースケッチである.

図 6: LHQ データセットでのスケッチ 2 の人が描いたカラースケッチからの生成例. Color Sketch は入力となるペイントソフトを用いて人が描いたカラースケッチある.

深度マップが質感の向上に影響するほどベースの pix2pixHD の精度が良くなかったためであると考えられる.

また, 各手法の LHQ データセットでの FID スコアと KID スコアは, ともに ADE20K データセットでの実験結果よりも高精度な結果が得られた. 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, LHQ データセットが自然風景のみのデータセットであり, ADE20K データセットのほう変換が困難であったためだと考えられる. さらに, LHQ データセットでの結果で人工物が上手く生成できない例が得られたが, これは, LHQ データセットに人工物を含む画像が少なかったためだと考えられる.

人が描いたスケッチからの生成結果については, 人が描いたスケッチからであってもフォトリアルに変換可能であることが確認できたが, 自動生成したスケッチからの生成と比較すると細かな質感の面で劣っていることが確認できた. これは, 自動生成により得られたスケッチのみを学習の入力としていたために, スケッチのスタイルに制約が生まれてしまっているためだと考えられる.

5 おわりに

本研究では, ラベルや言語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カラースケッチ画像を入力とする画像生成において, 生成手法として Diffusion Model に基づく U-Net と DPT-Large を導入した pix2pixHD を提案した. 実験結果より, Diffusion Model に基づく U-Net と DPT-Large による深度推定結果の活用はスケッチからの画像生成において精度の向上に効果的であることが確認できた. 本研究では, DPT-Large による深度推定結果の活用が精度の向上につながったが, 画像のセグメンテーションは深度推定だけでなく, セマンティックセグメンテーションが存在し, STEGO(Self-supervised Transformer with Energy-based

Graph Optimization) [28] や PiCIE(Pixel-level feature Clustering using Invariance and Equivariance) [29] のような教師なしセマンティックセグメンテーションモデルの活用についても検討する必要があると考えられる.

また, 人が描いたスケッチから生成が可能であることも確認したが, 自動生成したカラースケッチからの生成結果と比較すると細かな質感の面で劣っており, より多様なスタイルのスケッチや粗いスケッチには対応することが出来ないと考えられる. 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が, より粗くスタイルが異なっていてもフォトリアルに変換可能な手法が必要であると考えられる.

今後の課題としては, Diffusion Model を使用し SDEdit [15] を導入することでスケッチのスタイルへの依存を軽減することや ControlNet を導入し, 学習済み Diffusion Model を用いることによる精度改善が考えられる.

文　献

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

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[3] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[5] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021.

[6] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

[8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[10] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

[11] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[12] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[13] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[14] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. In *arXiv*, 2022.

[15] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[17] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[18] C. Y. Suen T. Y. Zhang. A fast parallel algorithm for thinning digital patterns. In *Communications of the ACM*, volume 27, pages 236–239, 1984.

[19] Michael Van den Bergh, Xavier Boix, Gemma Roig, and Luc Van Gool. SEEDS: superpixels extracted via energy-driven sampling. *CoRR*, abs/1309.3848, 2013.

[20] Simon Niklaus. A reimplementation of HED using PyTorch, 2018.

[21] Clark Alex. Pillow (pil fork) documentation, 2015.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[23] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[24] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14144–14153, October 2021.

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[26] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[27] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

[28] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022.

[29] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16794–16804, June 2021.