Vol.2018-NL-238 No.15
2018/12/12

テキストに基づく単一オブジェクト画像生成における描画の高品質化

野本 英梨子[1,a)]　Chenhui Chu[2,b)]　荒瀬 由紀[1,c)]

概要：テキストに基づく画像生成は，入力されたテキストに記述された通りの画像を生成することを目的とする．生成された画像は文の意味を理解する手助けとして言語学習者支援やコミュニケーション支援に利用できると期待されている．単一のオブジェクトに言及するテキストに基づく画像生成は高品質に行えるが，複数のオブジェクトに言及するテキストに基づく画像生成ではオブジェクトが正確に描画されないという課題がある．本研究では複数のオブジェクトに言及するテキストに基づく画像生成を高品質に行うことを目指し，以下のようなフレームワークを提案する．まずテキストに基づいて全景画像を生成したのち，描画されるべきオブジェクトをテキストから特定しそのオブジェクト部分のみを全景画像から切り出し高品質化する．そして高品質化したオブジェクト画像を元の全景画像に統合する．本稿ではフレームワークの有効性を検証するため，既存研究で取り組まれてきた単一のオブジェクトに言及するテキストに基づく画像生成の高品質化に取り組んだ．自動評価および人手評価により，オブジェクト高品質化手法が有効であることを示した．また，実験結果の詳細な分析により高品質化においては元の画像の有用な特徴を保持する必要があること，テキストの内容をより強く反映する必要があること，また高品質化すべき画像の事前選別が必要であることを示した．

Eriko Nomoto[1,a)]　Chenhui Chu[2,b)]　Yuki Arase[1,c)]

1. はじめに

画像生成技術は画像変換や画像修復など様々な分野で使用される．Generative Adversarial Networks（GAN）[1]の台頭により高品質な画像生成が可能になった．GAN による画像生成研究の主流はランダムなノイズから画像を生成しているが [2], [3]，テキストを入力とし，入力テキストに記述された通りの画像を生成するタスクも取り組まれ始めている [4], [5], [6], [7], [8], [9], [10]．これが可能になると，生成画像は文の意味を理解する手助けになり，言語学習者や異なる言語話者間のコミュニケーション支援に利用できると期待される．

テキストに基づく画像生成において，学習データセットを鳥や花など特定のドメインに絞った場合は既存手法 [6] を用いて高品質な画像を生成できる．しかしより一般的に，被写体や場面に制約がなく，画像内の複数のオブジェクトが言及されるようなデータセットに適用すると，十分な品質の画像が生成されないことが Hong ら [9] によって示されている．背景などの大まかな色合いには問題がない一方で，物体（オブジェクト）の形や詳細な部分を正確に描画できていないことが多い．

テキストに基づく高品質な複数オブジェクト画像生成を目指し，本研究では既存手法によって生成した画像に含まれるオブジェクトごとに，詳細部分を正確に描画し高品質化するフレームワークを提案する．オブジェクト部分の高品質化は，全景からオブジェクト部分を抽出した部分画像とテキストを入力とし，高品質化した部分画像を GAN によって生成する．提案フレームワークの有効性を示すため，既存研究で取り組まれてきた単一オブジェクト生成タスクである鳥の画像データセットで実験を行った．その結果，オブジェクト高品質化は有効であることが示されたが，一方で高品質化が改悪につながる場合があり，高品質化において元の画像の有用な特徴を保持する必要があること，テキストの内容をより強く反映する必要があること，高品質

[1] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
[2] 大阪大学データビリティフロンティア機構
Institute for Datability Science, Osaka University
[a)] nomoto.eriko@ist.osaka-u.ac.jp
[b)] chu@ids.osaka-u.ac.jp
[c)] arase@ist.osaka-u.ac.jp

© 2018 Information Processing Society of Japan

図 1　提案フレームワーク概要

図 2　品質の低い生成画像の例

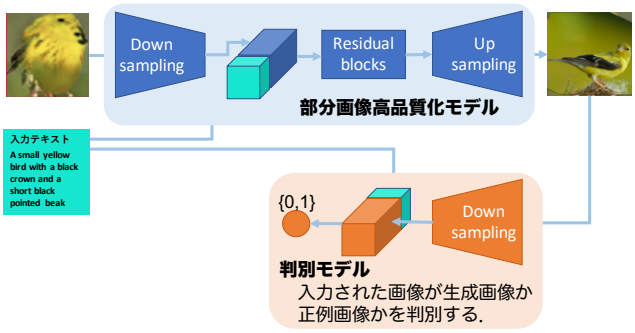


図 3　部分画像高品質化モデルの概要

化すべき画像の事前選別が必要であることを示した．

2. 提案フレームワーク

本研究で提案するフレームワークを 図 1 に示す．まず，Stacked Generative Adversarial Networks（StackGAN）[6] を用いて入力テキストから全景画像を生成する（2.1 節）．次に一般物体認識モデル [11] を用いて描画対象であるオブジェクトの位置を特定する（2.2 節）．次にこのオブジェクト部分のみを部分画像として切り出し，本研究で提案する部分画像高品質化モデルを用いて高品質化する（2.3 節）．最後に 高品質化した画像を元画像に自然に埋め込むため，Poisson Image Editing [12] を用いて得られた高品質化部分画像を全景画像に埋め込むことによって高品質化全景画像を取得する（2.4 節）．

2.1 全景生成

テキストに基づく全景の生成には StackGAN [6] を用いる．StackGAN による画像生成は 2 つの段階に分けられる．第 1 段階ではテキストに基づいて 64×64 ピクセルの低解像度カラー画像が生成される．第 2 段階ではテキストおよび生成された低解像度画像に基づいて 256×256 ピクセルの高解像度カラー画像が生成される．それぞれの段階で独立した GAN が画像生成に用いられる．

2.2 部分画像抽出

訓練済み一般物体認識モデル [11] を用いて，生成した全景画像内のオブジェクト部分のみを切り出す．一般物体認識モデルは入力画像内のオブジェクト位置を示す矩形と，それぞれの矩形部分のオブジェクトカテゴリを予測する．図 2 のように生成画像が自然に撮影された写真のような画像と大きく異なる場合，生成画像中から正しくオブジェクトを認識することは困難であることが予想される．提案手法ではこの問題を避けるために，生成画像中に描画されるべきオブジェクトのカテゴリをテキスト情報から事前に予想し，一般物体認識モデルの出力のうち該当カテゴリの尤度がもっとも高い矩形を選択する．具体的には，入力テキストに含まれる名詞を hypernym, hyponym に含むオブジェクトラベルをオブジェクトのカテゴリとする．図 1 の例では，“a small yellow bird” という記述からオブジェクトのカテゴリとして “BIRD” が抽出される．

2.3 部分画像高品質化

部分画像高品質化モデルは 2.2 節によって得られた部分画像および入力テキストを元に，同サイズでより高品質な画像を生成する．StackGAN の第 2 段階と同様に部分画像高品質化モデルと判別モデルを敵対的に訓練する．部分画像高品質化モデルの概要を 図 3 に示す．判別モデルは入力された画像が正例画像なのか部分画像高品質化モデルによる生成画像なのかを判別するモデルである．これに対し部分画像高品質化モデルは，生成画像が判別モデルに正例画像だと認識されるように，リアルな画像の生成を目指す．判別モデルと部分画像高品質化モデルの最適化を交互に行うことによって，部分画像高品質化の学習を行う．

部分画像高品質化モデルおよび判別モデルの訓練はオブジェクトカテゴリごとに行う．例えば対象のオブジェクトが “BIRD” カテゴリを持つのであれば，“BIRD” カテゴリ専用の部分画像高品質化モデルを使用する．複数オブジェクト画像の高品質化の際には，テキストから高品質化対象のオブジェクトに関する記述を抽出し，対応する高品質化モデルの入力とする．本研究では単一オブジェクト画像を対象としているため，テキスト全文を入力する．

部分画像高品質化は全景生成で大まかに描画した情報に

© 2018 Information Processing Society of Japan

従いオブジェクトの詳細を描画することを担うため，本研究では部分画像高品質化モデルによる生成画像は以下の 3 つの要件を見たす必要があると定義する．

(1) 部分画像の有用な特徴（オブジェクトの形態・向き・位置）を維持していること
(2) 部分画像の周辺と全景画像が連続的に繋がっていること
(3) オブジェクトが正例画像のようにリアルに描画されていること

本研究では要件 1 および要件 3 の実現を目指した平均画像損失と要件 2 および要件 3 の実現を目指した境界色調差損失の 2 種類の損失関数を設計し部分画像高品質化モデルに適用する．[*1]

2.3.1 平均画像損失

平均画像損失を

$$\begin{aligned}\mathcal{L}_{G_{\text{mean}}} &= E_{(x,t)\in p_{\text{data}}}[\log(1-D(G(x,\varphi_t),\varphi_t))] \\ &\quad + E_{(x,t)\in p_{\text{data}}}[\log(1-D(\frac{G(x,\varphi_t)+x}{2},\varphi_t))] \\ &\quad + \lambda D_{\text{KL}}(\mathcal{N}(\mu(\varphi_t),\Sigma(\varphi_t))||\mathcal{N}(0,1)) \qquad (1)\end{aligned}$$

と定義する．第 1 項は部分画像高品質化モデルの生成画像を判別モデルに入力した結果を用いた損失であり，生成画像が満たすべき要件 3「オブジェクトが正例画像のようにリアルに描画されていること」を担う．第 2 項は部分画像高品質化モデルへの入力画像と生成画像とを平均した画像を判別モデルに入力した結果を用いた損失であり，生成画像が満たすべき要件 1「部分画像の有用な特徴を維持していること」を担う．第 3 項は StackGAN で用いられている正則化項である．

p_{data} は訓練データセットであり $x, \hat{x} \in \mathcal{N}^{H\times W\times 3}$ はそれぞれ高さ H，幅 W，チャネル数 3 を持つ部分画像，正例画像を表すテンソルである．φ_t はテキスト t の分散表現である．また，G，D はそれぞれ部分画像高品質化モデル，判別モデルである．$D_{\text{KL}}(\cdot||\cdot)$ は 2 つの確率分布間の KL 分散であり，$\mathcal{N}(\mu,\Sigma)$ は平均 μ，分散 Σ の正規分布である．$\mu(\varphi_t)$ や $\Sigma(\varphi_t)$ は φ_t から得られるベクトルで，モデルの中で訓練される値である．$\lambda \in R$ は正則化項の重みを決定するハイパーパラメータである．

2.3.2 境界色調差損失

境界色調差損失を

$$\begin{aligned}\mathcal{L}_{G_{\text{color}}} &= E_{(x,t)\in p_{\text{data}}}[\log(1-D(G(x,\varphi_t),\varphi_t))] \\ &\quad - \sum \sigma(F(x))\log\sigma(F(G(x,\varphi_t))) \\ &\quad + \lambda D_{\text{KL}}(\mathcal{N}(\mu(\varphi_t),\Sigma(\varphi_t))||\mathcal{N}(0,1)), \qquad (2) \\ F(x) &= \frac{\sum_{q\in S} x_q}{||S||} \qquad (3)\end{aligned}$$

と定義する．式 (3) は画像の周囲 ω ピクセルの RGB それぞれの平均を表す．第 2 項は部分画像高品質化モデルへの入力画像と生成画像それぞれの色調のシグモイドクロスエントロピー誤差であり，生成画像が満たすべき要件 2「部分画像の周辺と連続的に繋がっていること」を色調の観点から担う．σ はシグモイド関数，S は画像の周囲 ω ピクセルを満たす座標集合であり，$x_q \in R^3$ である．

[*1] 全ての要件を満たす損失関数の設計は今後の課題とする．

図 4 全景への統合例．全景の該当部分を高品質化した部分画像で置き換えた場合（左）と，Poisson Image Editing[12] を用いて高品質化下部分と背景が滑らかに繋がるように埋め込んだ場合（右）

2.3.3 判別モデル損失

判別モデルの損失関数は StackGAN と同じ式 (4) を用いる．

$$\begin{aligned}\mathcal{L}_D &= E_{(\hat{x},t)\in p_{\text{data}}}[\log D(\hat{x},\varphi_t)] + \\ &\quad E_{(x,t)\in p_{\text{data}}}[\log(1-D(G(x,\varphi_t),\varphi_t))] \qquad (4)\end{aligned}$$

2.4 全景への統合

高品質化した部分画像を全景へ統合する．その際，高品質化によって背景の色調などが変わりうるため，そのままピクセルを置き換えると **図 4** 左に示すように境界付近に違和感を生じてしまうことがある．背景と滑らかにつながった部分画像の埋め込みを行うために，Poisson Image Editing [12] を用いる．Poisson Image Editing では，画像埋め込みを全景における部分画像位置を穴とみて，穴の内側と外側との勾配が小さくなりかつ内側の勾配が高品質化部分画像の勾配に近づくようにこの穴を埋める問題として定式化する．これにより高品質化画像を全景画像に滑らかに埋め込むことができる．

3. 評価実験

提案フレームワークを検証の有効性するため，評価実験を実施した．本実験では部分画像高品質化手法の性能を評価するため，既存研究で取り組まれてきた単一オブジェクトの生成画像を対象とする．実験手順を **図 5** に示す．実験用のデータセットとして CUB [13] を用いた． CUB はもともと画像キャプション生成用のデータセットであり，画像 1 枚に対し平均 5 件の説明文が与えられている．また，画像中のどこにオブジェクトが写っているかを表す矩形が与えられている．本実験ではこの画像のオブジェクト

© 2018 Information Processing Society of Japan

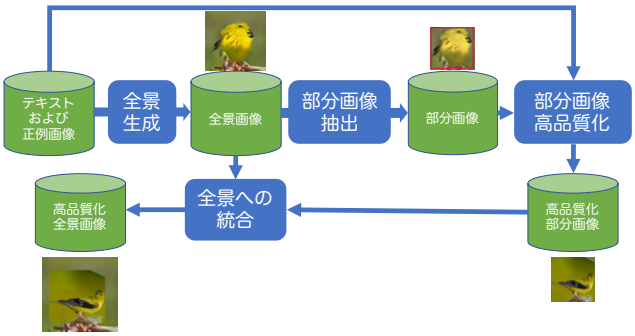


図 5 実験手順の概要．データセットのテキストを用いて全景を生成し，部分画像を抽出．訓練セットのテキストと生成した部分画像を用いて部分画像高品質化モデルを訓練する． 開発セットでハイパーパラメータのチューニングを行ったのち，テストセットを用いて高品質化部分画像とそれを全景へ統合した画像の評価を行う．

表 1 全景生成画像および部分画像抽出の品質

品質の分類	全 118 件中の件数（割合 [%]）
生成・抽出共に問題なし	66 件（55.9%）
生成失敗	15 件（12.7%）
抽出失敗	37 件（31.4%）

部分を切り抜いて，部分画像高品質化モデルを訓練する際の正例として用いた．訓練セットのテキスト 88,550 文に対し，1 文につき 1 枚の全景画像を生成した． 開発セット・テストセットのテキスト 29,330 文に対し開発セットは 1 文につき 2 枚，テストセットは 1 文につき 1 枚の全景画像を生成した． それぞれの全景画像に対し，オブジェクトカテゴリとして “BIRD” を持つ部分画像の抽出を行った．以上の操作により訓練セット 88,550 枚， 開発セット 58,660 枚，テストセット 29,330 枚の全景画像および部分画像を得た．

3.1 全景画像生成および部分画像抽出の評価

基礎となる StackGAN における全景画像生成の品質および部分画像抽出の性能を検証するため，実験を行った．テストセットの全景画像および部分画像を 118 対ランダムにサンプリングし，主著者 1 名により「生成失敗」,「抽出失敗」,「生成・抽出共に成功」の 3 つのグループに分類した．分類の基準を以下に示す．

生成・抽出共に問題なし
鳥らしきものが描画されており，その部分が部分画像として概ね正しく抽出されている．

生成失敗
そもそも鳥らしきものが描画されていない．

抽出失敗
鳥らしきものが描画されているが，鳥の部分画像抽出に失敗している．

分類の結果を **表 1** に示す．「生成・抽出共に成功」の画像は高品質化を効果的に行うことができると期待され，このような画像は 66 件（55.9%）であった．

「生成失敗」や「抽出失敗」に分類されるデータは自動的に検知し，取り除く必要がある．特に全景生成に用いたモデルはある 1 文に対し任意の枚数の画像を生成できるため，全景や部分画像抽出の品質が低いものを検出する方法があれば，複数枚の画像を生成しその中で品質の高いものを用いることができる．このような低品質な生成画像の検出手法については今後の課題とし，本実験では全ての画像を評価に用いる．

表 2 各ハイパーパラメータの設定

パラメータ名	パラメータの値
H	256
W	256
λ	1
ω	16

表 3 Inception Score

	Inseption Score	
画像種別	部分画像	全景画像
高品質化前	3.51 ± .02	3.55 ± .04
平均画像損失モデル	3.68 ± .04	3.59 ± .04
境界色調差損失モデル	3.76 ± .03	3.68 ± .04

3.2 自動評価指標

生成画像の自動評価指標として Inception Score [14] を用いた．Inception Score は画像生成モデルによる生成画像の品質と多様性を評価する指標であり，画像生成研究で広く用いられている．Inception Score は以下の式で定義される．

$$I = \exp(E_x[D_{\mathrm{KL}}(p(y|x)||p(y))]). \tag{5}$$

x はモデルによる生成画像であり，y は画像認識モデルである Inception Model [15] によって予測されるラベルである．この指標は，良い生成モデルほど意味のある画像を生成し画像認識モデルに特定のラベルを強く予想させ，また多様な画像を生成するため画像ごとに画像認識モデルの予測ラベルは様々であろうという直感に基づいて設計されている．Inception Score は定義上多くの画像によって算出される必要があり，本研究では 開発セット 58,660 枚，テストセット 29,330 枚の高品質化画像を元に算出した．先行研究 [6] に従い Inception Model は CUB データを用いてファインチューニングしたもの[*2] を使用した．

3.3 部分画像高品質化時のパラメータ設定

部分画像高品質化モデルのハイパーパラメータの設定を **表 2** に示す．開発セットにおける Inception Score が最大となるようにこれらハイパーパラメータを調整した．

[*2] https://github.com/hanzhanggit/StackGAN-inception-model

© 2018 Information Processing Society of Japan

表 4　A〜E の画像の平均順位および 95% 信頼区間

データ種別		A	B	C	D	E
鳥の色を	部分画像	1.09 ± 0.87	2.90 ± 2.22	3.55 ± 1.94	3.78 ± 1.93	3.68 ± 2.43
評価した場合	全景画像	1.12 ± 0.97	2.74 ± 2.08	3.47 ± 1.95	3.95 ± 1.91	3.73 ± 2.36
鳥の形を	部分画像	1.19 ± 1.23	2.53 ± 2.05	3.63 ± 2.17	3.86 ± 1.79	3.79 ± 2.16
評価した場合	全景画像	1.11 ± 0.84	2.49 ± 1.91	3.57 ± 2.05	3.96 ± 1.72	3.88 ± 2.09

表 5　鳥の色を評価した場合に
高品質化の前後で順位が向上した画像の割合

	順位の増加割合 [%]	
画像種別	部分画像	全景画像
平均画像損失モデル	41.75	36.25
境界色調差損失モデル	46.75	42.25

表 6　鳥の形を評価した場合に
高品質化の前後で順位が向上した画像の割合

	順位の増加割合 [%]	
画像種別	部分画像	全景画像
平均画像損失モデル	43.00	38.50
境界色調差損失モデル	46.75	43.00

3.4　生成画像の評価

生成画像の評価は Inception Score による自動評価と，同一のテキストから提案手法や比較手法によって生成した画像を人手で順位付けする人手評価を行った．これは Inception Score ではテキストと生成画像の整合性を評価できないためである．

3.4.1　Inception Score による自動評価

高品質化前の画像，平均画像損失モデルによる高品質化画像，境界色調差損失モデルによる高品質化画像それぞれについて，部分画像および部分画像を全景に統合した全景画像の Inceptino Score を **表 3** に示す．高品質化によって部分画像における Inception Score は 4.84〜7.12% 上昇しており，高品質化に成功していると言える．一方でこの部分画像を全景に統合して取得した全景画像においては Inception Score が 1.13〜3.66% の上昇に止まっている．これは，部分画像高品質化の際に背景などの特徴を大きく変えてしまい，全景との整合性が保たれなかったためと考えられる．

3.4.2　人手評価

Inception Score は画像のみに基づく評価指標であるため，入力テキストと生成画像との整合性を評価できない．そこで以下に示す人手評価を全景画像，部分画像それぞれについて行なった．評価セットから 400 対のテキストと正例画像をランダムに選択し，それぞれのテキストを用いて以下の A〜E の 5 種類の画像を取得もしくは作成した．

A. 描画元のテキストに紐づいている正解画像
B. 正解画像の中で A の画像とは別のテキストに紐づいている画像からランダムに選んだ画像
C. テキストから [6] を用いて生成した画像
D. 平均画像損失モデルを用いて C を高品質化した画像
E. 境界色調差損失モデルを用いて C を高品質化した画像

A〜E の画像を 1 セットとし，全 400 セットを 50 セットずつに 8 分割し， Amazon Mechanical Turk のワーカ 80 名にランダムに振り分け，「鳥の色がテキストに合致しているか」「鳥の形がテキストに合致しているか」のそれぞれの観点から A〜E の画像の順位付けを行ってもらった．その際，テキストを読まずに順位付けを行う質の低いワーカを除外するため，A よりも B の順位が高いような回答が一定数以上存在するワーカを不採用とした．

順位付けの結果，50 セットずつの各サブセットに対し 5〜15 名のワーカによる順位が得られた．この順位セットから CroudOrdering [16] を用いて真の順位を推定した．ワーカの一致率を示すスピアマンの順位相関係数は部分画像において平均 0.41，標準偏差 0.13，全景画像においては平均 0.41，標準偏差 0.12 であった．A〜E の画像の平均順位とその 95% 信頼区間を **表 4** に示す．これより，高品質化前後で顕著な順位の変動は見られなかったが，一貫して境界色調差損失モデルの方が平均画像損失モデルより高い平均順位となった．

さらに詳細を検証するため，高品質化前と後の順位の増減を集計した．平均画像損失モデルおよび境界色調差損失モデルのそれぞれについて，高品質化前の画像よりも高品質化後の画像の方が順位が向上した割合を **表 5**，**表 6** にそれぞれ示す．部分画像においては 41.75〜46.75% のサンプルが高品質化によって順位が向上している．一方，全景画像において部分画像高品質化の結果順位が向上したサンプルは 36.25〜43.00% となった．これは Inception Score による自動評価同様，部分画像高品質化時に全景との整合性が保たれなかったためと考えられる．

3.4.3　入力テキストの影響

本項では入力テキストが高品質化に与える影響について分析する．入力テキストの長さの高品質化への影響を調べるため，高品質化に成功した（高品質化後の順位が向上した）セットと失敗した（高品質化後の順位が低下した）セットのそれぞれについて，入力テキストの単語数別の度数分布を **図 6**，**図 7** に示す．これによると，鳥の色について評価した場合と鳥の形について評価した場合のどちらも，テキスト長が短い方がやや高品質化に成功しやすいと言える．これは入力テキストが長くなるほど画像の描画すべき特徴が詳細に記述されるため，そのような詳細な特徴を再

© 2018 Information Processing Society of Japan

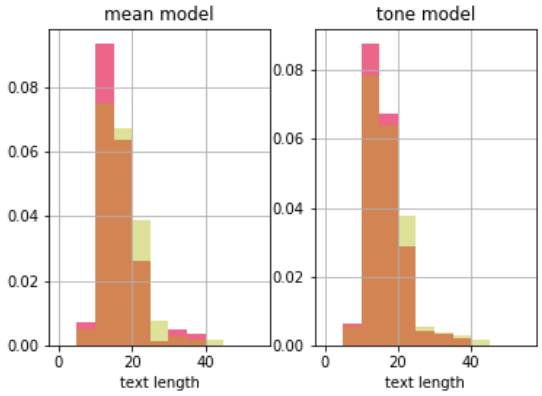


図 6 入力テキストの単語数別の高品質化成功，失敗の割合．平均画像損失モデル（mean model） および境界色調差損失モデル（tone model） それぞれについて，部分画像において鳥の色を評価した場合に高品質化に成功した割合（赤色）および失敗した割合（黄色）．

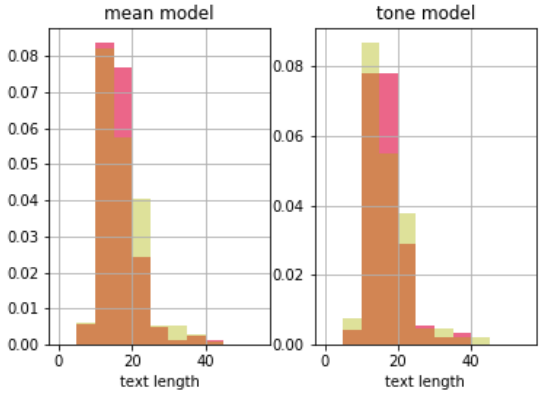


図 7 入力テキストの単語数別の高品質化成功，失敗の割合．平均画像損失モデル（mean model） および境界色調差損失モデル（tone model） それぞれについて，部分画像において鳥の形を評価した場合に高品質化に成功した頻度分布（赤色）および失敗した頻度分布（黄色）．

現する高品質化が困難であるためと考えられる．

次に，入力テキストに含まれている語彙の高品質化への影響を調べるため，入力テキストに単語 w が現れる場合の高品質化結果の条件付き確率 $P(\text{success}|w)$ （高品質化後に順位が向上した場合）または $P(\text{fail}|w)$ （高品質化後に順位が低下した場合）の上位 20 単語（名刺，動詞，形容詞，副詞）を **表 7**，**表 8** に示す．

表 7 によると，鳥の色について評価した場合，平均画像損失モデルでは色に関する単語が高品質化後に順位が向上したケースでは 6 件，順位が向上しなかったケースでも 6 件現れている．それぞれの語彙をみると，“green” や “grey” など暗い色の生成には成功しているが，“yellow” や “red” など彩度の高い色が必要な画像の生成に失敗しやすいことがわかる．一方，境界色調差損失モデルでは色に関する語彙が高品質化後に順位が向上したケースで 1 件，順位が向上しなかったケースで 4 件しか現れていない．このことから，境界色調差損失モデルのテキストに現れる色に対する再現能力は限定的と考えられる．

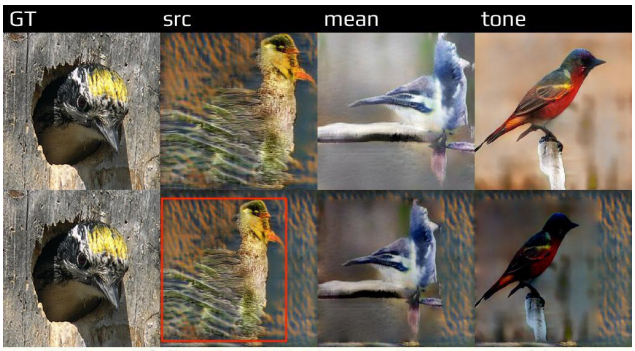
表 7 鳥の色を評価した場合の入力テキスト中の単語と高品質化結果の条件付き確率の上位 20 単語．色に関する単語を太字で表す．

平均画像損失モデル		境界色調差損失モデル	
成功	失敗	成功	失敗
thick	rest	thick	bars
water	**colorful**	face	little
tiny	tail	sized	along
cheek	hooked	spots	coverts
pointy	crest	legs	tarsus
chest	body	speckled	mostly
sized	**dark**	thin	**yellow**
green	mostly	water	throat
darker	nape	chest	back
back	**yellow**	cheek	breast
eyering	pointed	medium	wings
medium	small	sharp	**darker**
secondaries	**bright**	around	head
tarsus	**red**	eyering	crown
primaries	**brown**	curved	**blue**
grey	feet	tiny	**dark**
orange	compared	tail	small
tan	feathers	top	feet
gray	eyes	**red**	belly
stripe	wings	neck	light

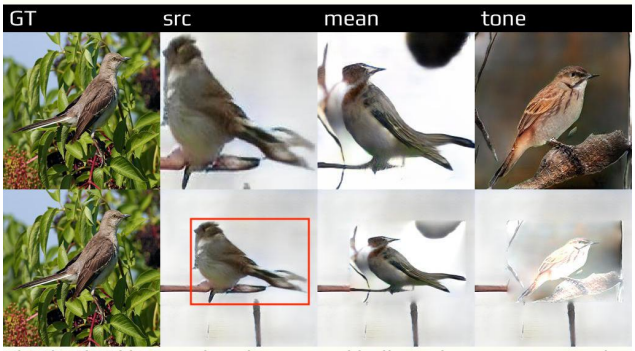
表 8 鳥の形を評価した場合の入力テキスト中の単語と高品質化結果の条件付き確率の上位 20 単語．鳥の形や模様に関する形容詞を太字で表す．

平均画像損失モデル		境界色調差損失モデル	
成功	失敗	成功	失敗
water	legs	**rounded**	bars
compared	**hooked**	**spots**	**little**
darker	dark	**thin**	**hooked**
eyering	blue	water	dark
body	**wide**	compared	light
chest	**thick**	legs	wing
sized	coverts	red	throat
sharp	cheek	**large**	back
throat	eyes	**tipped**	**small**
primaries	gray	**sized**	belly
eye	**little**	primaries	crown
feet	**pointed**	**patch**	top
grey	wingbars	**colorful**	yellow
nape	**short**	**long**	blue
tan	tail	**pointy**	wings
curved	along	**curved**	gray
stripe	bars	**tiny**	mostly
pointy	neck	neck	**short**
mostly	belly	**speckled**	grey
wing	colored	tail	head

© 2018 Information Processing Society of Japan



(a) this beautifully colored bird has a yellow crown, a black beak and eyes, and a white throat.



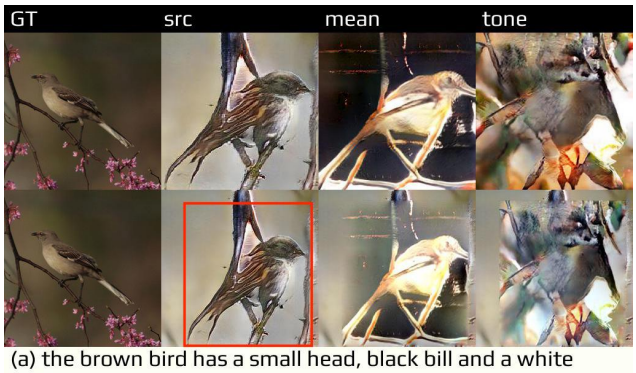
(b) this bird has a white breast and belly with gray crown and neck, with darker brown wings and inner reticles.

図 8 平均画像損失モデル，境界色調差損失モデルそれぞれにより高品質化に成功した例．上から順に部分画像，全景画像，入力テキストである．画像は左から順に正例画像（GT），高品質化前の画像（src），平均画像損失モデルによる高品質化画像（mean），境界色調差損失モデルによる高品質化画像（tone）である．高品質化前の全景画像には，部分画像として抽出された位置を赤枠で示す．

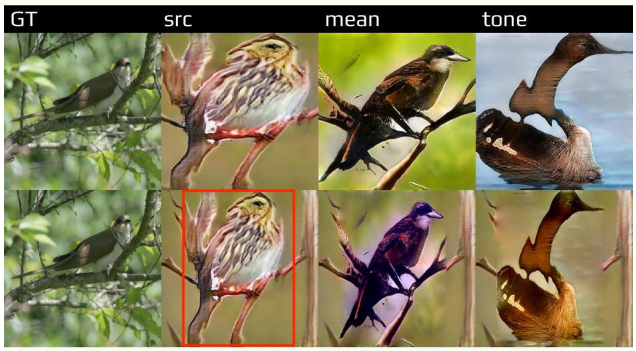
表 8 によると，鳥の形について評価した場合，境界色調差損失モデルでの高品質化に成功した語彙には鳥の形や模様に関する形容詞が 13 件現れており，平均画像損失モデルよりも極端に多い．このことは境界色調差損失モデルがテキストに記述されている細かい特徴の描写に強いことを意味する．これは色調差損失モデルが平均画像損失モデルに比べて入力画像（StackGAN が生成する鳥の描画）に左右されにくく，大まかな特徴しか捉えていない入力画像から細部の特徴を捉えた高品質化画像を生成しているためと考えられる．

3.4.4 定性的分析

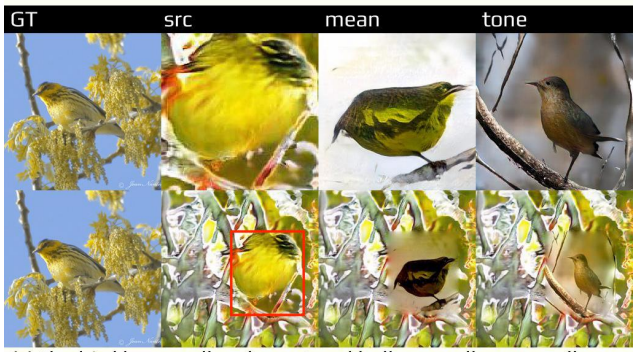
平均画像損失モデル，境界色調差損失モデルそれぞれによる高品質化が成功した（高品質化後に順位が順位が向上した）例を **図 8** に示す．図 8 (a) および (b) の両方で，部分画像においては高品質化前の画像よりも高品質化後の画像の方が鳥らしい画像となっている．平均画像損失モデルによる高品質化画像と境界色調差損失モデルによる高品質化画像を見比べると，特に 図 8 (a) において境界色調差損失モデルの方がよりリアルな鳥が描画されている．表 3 に示す自動評価，表 5 および 表 6 に示す人手評価の結果を



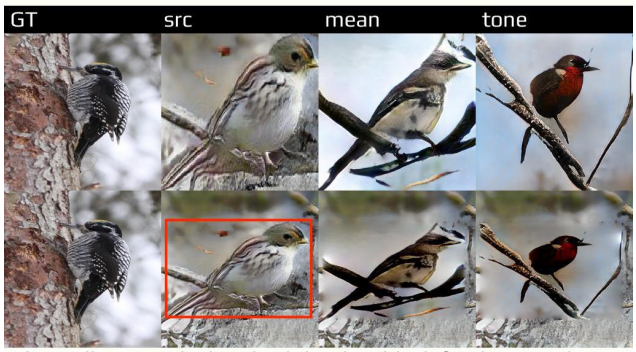
(a) the brown bird has a small head, black bill and a white belly.



(b) small white breasted bird, with brown wings , crown, and tail feather, brown beak, orange-brown eyes.



(c) the bird has a yellow breast and belly as well as a small bill.



(d) small grey and white bird that has black feathers on the side to its tail and a small yellow circle on its head.

図 9 平均画像損失モデル，境界色調差損失モデルそれぞれによる高品質化画像が高品質化前の画像よりも下位である例．上から順に部分画像，全景画像，入力テキストである．画像は左から順に正例画像（GT），高品質化前の画像（src），平均画像損失モデルによる高品質化画像（mean），境界色調差損失モデルによる高品質化画像（tone）である．高品質化前の全景画像には，部分画像として抽出された位置を赤枠で示す．

© 2018 Information Processing Society of Japan

見ても，境界色調差損失モデルの方が一般的に平均画像損失モデルより高品質化性能が高いことがわかる．一方で平均画像損失モデルによる高品質化で特に図 8 (b) において，高品質化前の鳥や枝などの輪郭が反映されており，全景と統合した際の整合性が境界色調差損失モデルに勝ることが見て取れる．図 8 (b) のように，背景が特徴的な場合では平均画像損失モデルの方が整合性の高い画像を生成できるため，これら 2 つのモデルの統合が今後の課題である．

平均画像損失モデル，境界色調差損失モデルそれぞれによる高品質化に失敗した（高品質化後に順位が順位が低下した）例を **図 9** に示す．図 9 (a) は，特に境界色調差損失モデルによる生成画像において顕著なように，高品質化モデルが鳥らしい画像を生成できなかった例である．このような挙動は GAN の学習の不安定さに起因し，StackGAN なども同様の挙動を示す [6]．図 9 (b) および (c) では StackGAN による生成画像の質が低いにも関わらず，高品質化によってより鳥らしいオブジェクトが描画されている．しかし，高品質化前の画像の有用な特徴を反映できておらず，テキストと整合しない描画を行ったために低い順位となったと考えられる．図 9 (b) の入力テキストによると，全体的に茶色く，胸が白い鳥を生成する必要があるが，高品質化後の鳥は胸まで茶色く，テキストと一致しない．このことから，部分画像高品質化モデルは高品質化前の有用な特徴を保持し，また高品質化画像生成時にテキスト内容をより反映するよう改善する必要がある．図 9 (d) は高品質化前の画像が充分にテキストの内容を描画したものであり，それ以上の高品質化が困難であった例である．このような画像はそもそも高品質化を行う必要がないため，高品質化が必要なものかそうでないかの判定を事前に行い，必要なものについてのみ高品質化を行うというようなアプローチで改善が見込めると考える．

4. 関連研究

GAN [1] が提案されて以降，GAN を用いた多くの画像生成技術が研究されてきた．その応用研究としてテキストに基づく画像生成研究も盛んに行われている [6], [8], [9], [10], [17]．[17] では画像生成に必要なテキストの分散表現の獲得に，同じクラスラベルを持つ画像分類モデルとテキスト分類モデルとのそれぞれの分散表現を対応させる手法 [18] を用いる．このようにして得られたテキスト分散表現は視覚的な特徴を有すると考えられ，これを Deep Convolutional Generative Adversarial Networks（DCGAN）[2] への入力とすることでテキストに基づく画像生成を行う．Stacked GAN（StackGAN）[6] はテキストに基づく画像生成問題を 全景の描画とオブジェクトの詳細な描画の 2 段階に分け，全景の描画を低解像度画像の生成，オブジェクトの詳細な描画を高解像度画像の生成とし，それぞれに GAN を訓練することで GAN の訓練の不安定さを軽減し，[17] の手法に比べて高解像度な画像の生成に成功している．Attentional Generative Adversarial Networks（AttnGAN）[8] は StackGAN にアテンション機構を組み込むことでテキストの内容を反映しようとし，StackGAN に比べてより高品質な画像の生成に成功している．

StackGAN はデータセットとして CUB [13] や Oxford-102 [19] を用い，鳥や花などの特定のオブジェクトのみを描画対象にした場合には高品質な画像生成が可能である．しかしより一般的で被写体や場面に制約のない説明文付き画像データセットである MSCOCO [20] に StackGAN を適用すると，生成される画像の多くは十分な品質ではなく [9]，背景の色味などは場面を反映してはいるものの，オブジェクトが歪んでいたり，背景と混ざり合ったりする．このような課題を解決すべく，低解像度から高解像度へ StackGAN よりも段階的に画像を生成することで学習を安定させる手法 [7], [10]，オブジェクトごとに大まかな位置を予測し，それを元に Semantic Layout を生成し着色することで画像を生成する手法 [9] など，様々な手法が提案されている．

本研究では StackGAN で生成した画像を入力とし，StackGAN と同様に，複数のオブジェクトに言及するテキストに基づく画像の生成問題を全景の描画とオブジェクトの詳細な描画に分けて取り組む．[9] ではオブジェクトごとにオブジェクトの形状を予測したのちに，そこに着色することで画像を生成するが，オブジュエクト形状の正確な予測は困難である．本研究ではオブジェクトごとに大まかに描画された画像をより詳細に描画することでオブジェクト部分の画像を生成する．

5. 結論

本研究では，複数のオブジェクトに言及するテキストに基づく画像生成に向けて，テキストから生成された全景をオブジェクトごとに高品質化するというフレームワークを提案した．また，オブジェクト部分の高品質化の有効性を検証するため，既存研究で取り組まれている単一のオブジェクトに言及するテキストに基づく画像生成において本フレームワークを適用した．その結果，部分画像を高品質化したもののうち約 40% が高品質化によって改善されており，オブジェクト部分の高品質化の有効性が示された．また実験結果の詳細な分析を実施し，高品質化において入力テキストが与える影響を分析した．また高品質化においては元の画像の有用な特徴を保持する必要があること，テキストの内容をより強く反映する必要があること，また高品質化すべき画像の事前選別が必要であることを示した．今後これらの課題に取り組む予定である．

© 2018 Information Processing Society of Japan

参考文献

[1] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative Adversarial Networks, *NIPS* (2014).
[2] Radford, A., Metz, L. and Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, *ICLR* (2016).
[3] Karras NVIDIA, T. and Aila NVIDIA Samuli Laine NVIDIA Jaakko Lehtinen, T.: Progressive Growing of GANs for Improved Quality, Stability, and Variation, *arXiv:1710.10196* (2017).
[4] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H.: Generative Adversarial Text to Image Synthesis, *ICML* (2016).
[5] Reed, S., Akata, Z., Mohan, S., Tenka, S., Schiele, B. and Lee, H.: Learning What and Where to Draw, *NIPS* (2016).
[6] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X. and Metaxas, D.: StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, *ICCV* (2017).
[7] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X. and Metaxas, D.: StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks, *arXiv:1710.10916*, (online), available from ⟨https://arxiv.org/pdf/1710.10916.pdf⟩.
[8] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X. and He, X.: AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks, *CVPR* (2018).
[9] Hong, S., Yang, D., Choi, J. and Lee, H.: Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis, *CVPR* (2018).
[10] Zhang, Z., Xie, Y. and Yang, L.: Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network, *CVPR* (2018).
[11] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S. and Murphy, K.: Speed/accuracy trade-offs for modern convolutional object detectors, *CVPR* (2017).
[12] Pérez, P., Gangnet, M. and Blake, A.: Poisson Image Editing, *SIGGRAPH*, pp. 313–318 (2003).
[13] Wah, C., Branson, S., Welinder, P., Perona, P. and Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset, Technical Report CNS-TR-2011-001, California Institute of Technology (2011).
[14] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X.: Improved Techniques for Training GANs, *NIPS* (2016).
[15] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: Rethinking the Inception Architecture for Computer Vision, *CVPR* (2016).
[16] Matsui, T., Baba, Y., Kamishima, T. and Kashima, H.: Crowdordering, Technical report.
[17] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H.: Generative Adversarial Text to Image Synthesis, *ICML* (2016).
[18] Reed, S., Akata, Z., Schiele, B. and Lee, H.: Learning Deep Representations of Fine-grained Visual Descriptions, *CVPR* (2016).
[19] Nilsback, M.-E. and Zisserman, A.: Automated flower classification over a large number of classes, *ICCVGIP* (2008).
[20] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollár, P.: Microsoft COCO: Common Objects in Context, *ECCV* (2014).

© 2018 Information Processing Society of Japan