人工画像を用いた Text-to-Image モデルの事前学習

中尾 純平 [1]　磯沼 大 [1,2]　片岡 裕雄 [3]　森 純一郎 [1,4]　坂田 一郎 [1]
[1] 東京大学　[2] エディンバラ大学　[3] 産業技術総合研究所　[4] 理研 AIP
jnmsww99@g.ecc.u-tokyo.ac.jp　{isonuma, isakata}@ipr-ctr.t.u-tokyo.ac.jp
hirokatsu.kataoka@aist.go.jp　mori@mi.u-tokyo.ac.jp

概要

近年，大規模 Text-to-Image (T2I) モデルの事前学習に用いるデータセットに対して倫理面の問題が指摘されている．そこで本研究では人工画像を用いた T2I モデルの事前学習を検討する．人工画像にはキャプションが付与されていないことが事前学習に用いる際の難所となるが，本研究では CLIP を活用して人工画像に疑似的なキャプションを付与することでこの点にアプローチした．評価実験では，人工画像と実画像でそれぞれ事前学習したモデルを，複数の実画像データセットでファインチューニングした後の生成画像の精度で比較することで，人工画像を用いた事前学習が実画像を用いた事前学習に迫る有効性を示すことと共に，人工画像の色による多様性と輪郭という性質が重要であることを確認した．

1　はじめに

近年，大規模 Text-to-Image (T2I) モデルの発展が目覚ましい．例えば，Stable Diffusion [1] は任意の入力テキストの内容に応じた画像を生成できるモデルとして脚光を浴びた．これらの T2I モデルは，LAION-5B [2] など数十億規模の画像及びキャプションがペアリングされたデータセットで事前学習を行い，その後特定のドメインのデータセットを用いてファインチューニングすることで構築される．大量の画像を用いた事前学習が，T2I モデルの高い汎化性能をもたらした．

しかし，LAION-5B のような大規模データセットは倫理面の問題が指摘されている．これらは主に web クロールで機械的に画像-キャプションペアを収集して構築される．従って，プライバシー侵害や著作権違反，公序良俗に反したコンテンツが含まれることが多い [3] が，その規模から人手による全数把握は極めて困難である．さらに，LAION-5B は欧州の個人データ保護規則である General Data Protection Regulation を侵害するとして，現在安全面での審査やデータ改善中と報告されている[1)]．

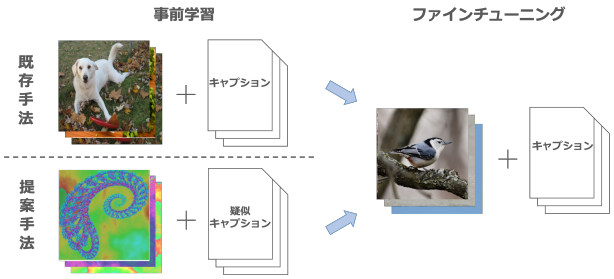


図 1: 提案手法の概要

近年，データの倫理問題解決に向けた人工画像を用いた事前学習が行われており，画像認識タスクでは実画像データセットの事前学習と比較してほぼ同等の精度が得られることが示されている [4, 5, 6]．この取り組みを踏まえて，本研究では，図 1 のように既存の T2I モデルが事前学習に用いている実画像を人工画像に置き換えることで，T2I モデルの構築が可能か検証する．人工画像は数式等のアルゴリズムで機械的に生成されるため，プライバシーや著作権，公序良俗に反するコンテンツを一切含まない．

T2I モデルの事前学習において，人工画像を用いる際の最大の難所は，人工画像に対応したキャプションの付与である．生成される人工画像にはキャプションが付与されておらず，事前学習に有効なキャプションも自明でない．そこで我々は，画像及びテキストの事前学習済みモデルである CLIP [7] の潜在空間を活用して実画像に対し疑似的なキャプションを付与する Lafite [8] に着目した．人工画像に対して，Lafite と同様に疑似キャプションを付与することで T2I モデルの事前学習を試みる．

評価実験では，人工画像で事前学習したモデルを複数の実画像データセットでファインチューニングし，生成された画像の評価を行った．事前学習を行わない場合に比べて，人工画像を用いた事前学習が

1) https://laion.ai/notes/laion-maintanence/

生成画像の精度と学習効率を高めることを確認すると共に，事前学習の文脈において倫理面で完全にクリアな T2I モデルを構築しつつ実画像で事前学習した場合に迫る精度が得られることを確認した．

2 提案手法

本研究では Lafite [8] と同様に，StyleGAN2 [9] をベースとした T2I モデルを用いる．ただし，提案する人工画像を用いた事前学習は diffusion model など他のモデルの事前学習にも利用できる．

本研究で使用するモデルは生成器と識別器で構成される．生成器はキャプションを入力として画像を生成する一方，識別器は入力画像が訓練画像か生成画像か判定する．生成器・識別器はそれぞれ次の損失関数 L'_G，L'_D を最小化するように学習される．

$$L'_G = L_G + \gamma L_{ConD} + \lambda L_{ConG} \quad (1)$$

$$L'_D = L_D + \gamma L_{ConD} \quad (2)$$

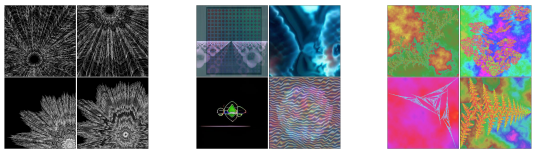
ただし，γ，λ は各項の重みを調整するハイパーパラメータである．L_G と L_D はいずれも GAN で一般に用いられる損失関数で，L_G は識別器が生成画像を訓練画像と誤識別するほど小さくなる一方，L_D は識別器が入力画像を正しく識別できるほど小さくなる．すなわち，生成器は識別器が訓練画像と誤識別する画像を生成するように，識別器は入力画像を正しく識別できるように学習する．また，L_{ConG} と L_{ConD} は生成器がキャプションを適切に反映した画像を生成するように設計された損失関数である[2]．

本研究では学習に用いられる画像を人工画像に，キャプションを疑似キャプションに置き換えて事前学習を行う．以下ではそれぞれについて説明する．

2.1 事前学習に用いる人工画像

本研究では，下記の 3 種類のアルゴリズムで生成された人工画像を用いる．近年，T2I モデルで生成された画像を事前学習に用いる取り組み [10] があるが，これらの画像もまたプライバシーや著作権などの問題が拭えない．従って本研究ではアルゴリズムで生成された人工画像のみを用いる．図 2 にこれらの人工画像の例を示す．

Visual Atoms [11] Kataoka ら [4] は数式で定義された関数により，多様な人工画像とそのクラスの機械的な生成を可能にする Formula-Driven Supervised Learning (FDSL) を考案した．Visual Atoms は FDSL の一種で，Visual Transformer (ViT) [12] の学習にオブジェクトの輪郭が重要である示唆 [5] から，正弦波の重ね合わせによる多様な輪郭表現を意図して設計された．Visual Atoms を用いた ViT の事前学習は，クラス分類やセグメンテーションにおいて，実画像を事前学習した場合とほぼ同等の精度を達成した．



(a) Visual Atoms (b) Shaders21k (c) Improved Frc.

図 2: 本研究で用いる人工画像の例

Shaders21k [6] Baradad らは既存の人工画像が単一アルゴリズムで生成されていることに着目して Shaders21k を考案した．Shaders21k は OpenGL から取得した 21,000 件のアルゴリズムで生成された多様な人工画像で構築される．Shaders21k は単一アルゴリズムで生成された人工画像と比較して，クラス分類で高い精度を達成している．

Improved FractalDB [13] Kataoka ら [4] は多様なグレースケールのフラクタル画像を含む FractalDB を考案した．しかし，FractalDB を構築するアルゴリズムには疎な画像が生成されうる等の問題点があった．Anderson らはこの問題点を解消したアルゴリズムを考案し，更に色の付与で多様性を高めたデータセット[3]を構築した．クラス分類では FractalDB と比較して高い精度を達成している．

2.2 疑似キャプションの付与

人工画像に疑似キャプションを付与する単純な手法として既存のキャプション生成モデルの利用が考えられる．しかし，モデルが人工画像に付与するキャプションは多様性が著しく低いことを実験的に確認している．例えば Visual Atoms には，その大半に “a black and white photo of an abstract design” といったキャプションが付与されてしまう．

Wang ら [14] は VQ-GAN [15] と CLIP [7] を活用して，画像のみで T2I モデルを学習する Clip-gen を考案した．しかし，Clip-gen は生成器が VQ-GAN 等のデコーダに制限される欠点がある．

2) 紙面の都合上，各損失の詳細については Zhou ら [8] を参照されたい．

3) 明確な名称がないため，本論文では Improved FractalDB と呼称する．

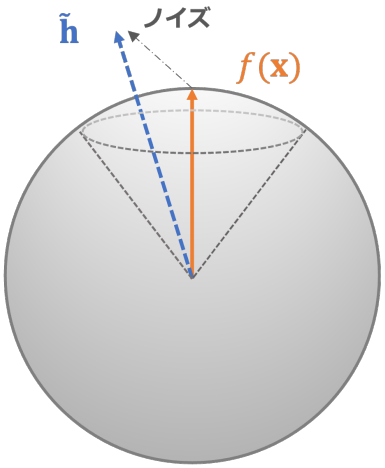


図 3: Lafite による疑似キャプション生成の概要図．点線で示した画像特徴量に近い領域に疑似キャプションが生成される．

そこで本研究は Zhou ら [8] の Lafite に着目した．Lafite では CLIP の潜在空間を活用して画像自体に疑似キャプションを表すベクトルを付与する．従って，モデルに対する依存性が小さく，本研究で用いるモデル以外でも利用可能である．

CLIP は画像とキャプションを共通した超球面上の潜在空間に埋め込む．このとき，ペアの画像とキャプションはコサイン類似度が大きくなるように埋め込まれる．従って，ある画像に対応したキャプションの文章特徴量は，共通潜在空間上の画像特徴量に近い領域に存在していると考えられる．この考察から Zhou らは画像特徴量にガウスノイズで摂動を加えて，対応したキャプションの文章特徴量の近似を検討した．具体的には画像を $\mathbf{x}$，CLIP の画像エンコーダを $f(\cdot)$，ガウスノイズを $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ とすると，画像 $\mathbf{x}$ に対する疑似キャプション $\mathbf{h}'$ は次式 (3) で計算される．

$$\mathbf{h}' = \frac{\tilde{\mathbf{h}}}{\|\tilde{\mathbf{h}}\|}, \quad \tilde{\mathbf{h}} = f(\mathbf{x}) + \xi\epsilon\frac{\|f(\mathbf{x})\|}{\|\epsilon\|} \tag{3}$$

ただし，$\|\cdot\|$ は L2 ノルムであり，ξ は摂動の強さを調整するハイパーパラメータである．図 3 に疑似キャプション生成の概要を示す．本研究では，Lafite を用いて人工画像に疑似キャプションを付与する．

3 実験

3.1 実験設定

事前学習 本論文では Visual atoms [11], Shaders21k [6], Improved FractalDB [13] について，それぞれ 100 万枚の画像を用いて前述のモデルを事前学習した．また，ベースラインとして事前学習を行っていないモデル（Scratch）と，広範なカテゴリの実画像で構成されたデータセットである Conceptual Captions (CC3M) [16] の内 100 万枚を用いて事前学習したモデルを用意した．この CC3M の事前学習では既存手法と同様にキャプションを用いて学習する設定 (CC3M+caption) と，疑似キャプションを用いて学習する設定 (CC3M+pseudo) を用意した．なお，事前学習時のエポック数はいずれのモデルにおいても 10 エポックで統一した．

ファインチューニング ファインチューニングに利用するデータセットとして，MSCOCO [17], CUB [18], CelebA-HQ [19] を用いる．MSCOCO は 91 種類の広範なカテゴリを含む画像で構成され，82,612 件を取得した．CUB は 200 種の鳥類を写した画像で構成され，8,849 件を取得した．CelebA-HQ は様々な人間の顔を写した画像で構成され，24,000 件を取得した．ファインチューニングでは各データセットからランダムにサンプルされた 10%の画像とそれに付与されたキャプションを用いる．

生成された画像の評価指標として，Inception Score (IS) [20] 及び Fréchet Inception Distance (FID) [21] を用いる．IS は生成された画像の品質と多様性を評価する指標で，高い値が望ましい．一方，FID は生成した画像とテストデータセットの画像の潜在空間上での分布間距離を表す指標で，低い値が望ましい．

実装の詳細については Appendix A.4 に記す．

3.2 生成された画像の精度評価

各事前学習モデルを CUB データセットでファインチューニングした場合について，学習ステップごとの FID の減少過程を図 4 に示す．実画像と人工画像のいずれにおいても，事前学習を行うことでファインチューニング後の生成画像の精度と学習効率が向上することがわかる．

ファインチューニング後の生成画像を複数のデータセット・指標で評価した結果を表 1 に示す．

人工画像 vs. Scratch 人工画像で事前学習したモデルは，事前学習を行わない場合（Scratch）と比較して CelebA-HQ の IS を除きいずれも高い精度を示した．人工画像での事前学習がファインチューニングでの精度を向上させることがわかる．

人工画像 vs. CC3M 人工画像を事前学習したモデルと比較して，実画像である CC3M で事前学

表 1: 生成画像の精度評価 (IS は大きい程良く，FID は小さい程良い)．上段は Scratch を除き実画像を，下段は人工画像を事前学習に使用．各段ごとに最もよい値を太字で表示．

	MSCOCO		CUB		CelebA-HQ	
	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓
Scratch	17.4	50.9	5.82	46.0	**3.02**	52.7
CC3M+caption	**25.9**	**20.4**	5.94	**20.4**	2.94	**17.8**
CC3M+pseudo	24.2	26.3	**6.21**	25.3	2.86	21.2
Visual Atoms	20.9	36.9	**6.95**	38.6	**2.85**	35.2
Shaders21k	22.1	37.4	6.64	35.3	2.67	**25.8**
Improved FractalDB	**24.2**	**29.3**	6.13	**30.4**	2.76	26.9

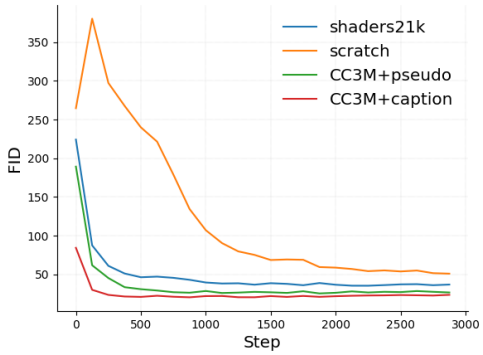


図 4: ファインチューニングにおける FID の減少過程（横軸は学習ステップ数）

表 2: 色を欠損させた場合の FID (小さい程良い)

	MSCOCO	CUB	CelebA-HQ
Improved FractalDB	29.3	30.4	26.9
w/o color	42.7	41.6	39.5

習したモデルは，いずれのデータセットでファインチューニングしても高い精度を示す．一方で，Improved FractalDB で事前学習したモデルは CC3M+pseudo に迫る精度を達成しており，実画像と比較しても人工画像の事前学習について一定の有効性が示された．

CC3M+caption vs. CC3M+pseudo　CC3M+caption と CC3M+pseudo を比較すると，CC3M+caption の方が精度が高いことがわかる．従って，疑似キャプションの利用が事前学習の有効性を低下させる可能性が示唆され，他の疑似キャプション付与方法の検討も今後の課題であると考えられる．

3.3　事前学習に適した人工画像の分析

3.2 節の結果より，事前学習の有効性が人工画像間で異なることがわかる．本節では，事前学習に適

表 3: 輪郭を欠損させた場合の FID (小さい程良い)

	MSCOCO	CUB	CelebA-HQ
w/o background	35.9	50.1	36.6
corrupted contour	41.6	52.7	41.3

した人工画像の性質を色と輪郭の観点で分析する．

色の重要性の分析　既存研究では，色による人工画像の多様性の向上が事前学習に重要であるという示唆がある [13]．そこで Improved FractalDB と，それをグレースケールに変換した画像 (w/o color) でそれぞれ 10 エポック事前学習したモデルの比較を行い色の重要性を検証する．表 2 に示す結果から，w/o color が Improved FractalDB に劣ることがわかる．従って，色による人工画像の多様性は事前学習に重要な性質と考えられる．

輪郭の重要性の分析　既存研究は人工画像の輪郭欠損が画像認識タスクにおける事前学習の有効性の低下を招くことを示した [5]．そこで Improved FractalDB の背景を黒色にした画像 (w/o background) と既存研究に倣い w/o background に黒色の直線をランダムな長さ・位置で 100 本描画して輪郭を欠損させた画像 (corrupted contour) でそれぞれ 10 エポック事前学習したモデルの比較を行い輪郭の重要性を検証する．表 3 に示す結果から，corrupted contour が w/o background に劣ることがわかる．従って，人工画像の輪郭は事前学習に重要な性質と考えられる．

4　おわりに

本研究では T2I モデルの人工画像を用いた事前学習を検討した．評価実験では人工画像での事前学習が実画像での事前学習に迫る有効性を示すことを確認した．また，色による多様性と輪郭という性質が人工画像に重要な性質である示唆を得て，T2I モデルの事前学習に用いる人工画像の方向性を示した．

謝辞

本研究は，NEDO JPNP20006 及び JST CREST JPMJCR21D1 の支援を受けたものである．

参考文献

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 10684–10695, 2022.

[2] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 25278–25294, 2022.

[3] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In **2021 IEEE Winter Conference on Applications of Computer Vision (WACV)**, pp. 1536–1546. IEEE, 2021.

[4] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images. In **Proceedings of the Asian Conference on Computer Vision**, 2020.

[5] Hirokatsu Kataoka, Ryo Hayamizu, Ryosuke Yamada, Kodai Nakashima, Sora Takashima, Xinyu Zhang, Edgar Josafat Martinez-Noriega, Nakamasa Inoue, and Rio Yokota. Replacing labeled real-image datasets with auto-generated contours. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 21232–21241, June 2022.

[6] Manel Baradad, Richard Chen, Jonas Wulff, Tongzhou Wang, Rogerio Feris, Antonio Torralba, and Phillip Isola. Procedural image programs for representation learning. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 6450–6462, 2022.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International conference on machine learning**, pp. 8748–8763. PMLR, 2021.

[8] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 17907–17917, 2022.

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 8110–8119, 2020.

[10] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. **arXiv preprint arXiv:2306.00984**, 2023.

[11] Sora Takashima, Ryo Hayamizu, Nakamasa Inoue, Hirokatsu Kataoka, and Rio Yokota. Visual atoms: Pre-training vision transformers with sinusoidal waves. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 18579–18588, 2023.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020.

[13] Connor Anderson and Ryan Farrell. Improving fractal pre-training. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision**, pp. 1300–1309, 2022.

[14] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. Clip-gen: Language-free training of a text-to-image generator with clip. **arXiv preprint arXiv:2203.00386**, 2022.

[15] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 12873–12883, 2021.

[16] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2556–2565, 2018.

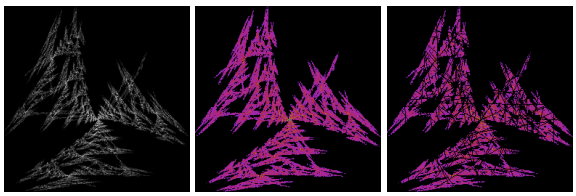
[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In **Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13**, pp. 740–755. Springer, 2014.

[18] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[19] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 2256–2265, 2021.

[20] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. **Advances in neural information processing systems**, Vol. 29, , 2016.

[21] 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. **Advances in neural information processing systems**, Vol. 30, , 2017.



(a) w/o color　(b) w/o background　(c) corrupted con.

図 5: 分析のため導入した人工画像の例

表 4: 各キャプション生成モデルで付与したキャプションの語彙数

	MSCOCO	Visual Atoms
BLIP	623	34
ClipCap	773	199

A　付録

A.1　キャプション生成モデルによるキャプション付与

2.2 節にて述べた，既存のキャプション生成モデルが人工画像に付与するキャプションの多様性が著しく低いことの分析結果を示す．具体的には，MSCOCO と Visual Atoms からランダムにサンプルした各 1000 件それぞれに BLIP と ClipCap でキャプションを付与し，それらのキャプションに含まれる語彙数を比較することで分析を行った．表 4 の結果から，いずれのモデルにおいても Visual Atoms に付与されたキャプションは MSCOCO に比べて語彙数が少ないことがわかり，キャプションの多様性の低さが確認される．

A.2　色・輪郭の重要性分析で用いた人工画像

3.3 節では，色と輪郭の性質が人工画像に重要な性質であることを示すために Improved FractalDB をグレースケールに変換した w/o color，Improved FractalDB の背景を黒色にした w/o background，w/o background の輪郭を欠損させた corrupted contour を導入した．これらの人工画像の例を図 5 に示す．

A.3　生成画像例

Scratch と CC3M+caption，CC3M+pseudo，Improved FractalDB で事前学習したモデルを，MSCOCO でファインチューニングした各モデルの生成画像例を図 6 に示す．

図 6: 各事前学習モデルの生成画像例

A.4　実装の詳細

損失関数と疑似キャプション生成時のハイパーパラメータは Lafite[4] の実装に従って設定した．損失関数のハイパーパラメータは $\gamma = 5$，$\lambda = 10$ で，疑似キャプション生成時の摂動の強さを調整するハイパーパラメータは $\xi = 0.25$ である．事前学習時の学習率は，生成器・識別器のいずれも 2.5×10^{-3} で統一している．一方，ファインチューニング時は MSCOCO に対して生成器は 2.5×10^{-3}，識別器は 1.0×10^{-3}，CUB・CelebA-HQ に対して生成器・識別器はいずれも 1.0×10^{-3} とした．最適化手法としては Adam を用いている．

ファインチューニング後の精度評価では，ファインチューニングに用いたデータセットと対応した検証データセットを評価に用いており，各モデルが FID の観点で収束した時点での結果を取得している．

4)　https://github.com/drboog/Lafite