I-004

CLIP を用いたテキスト駆動型の 3D 自然背景自動生成手法
Automatic Text-driven 3D Natural Background Generation Method Using CLIP

阿部 遥太朗† 渡邊 海斗† 中山弘敬‡ 白木厚司* 伊藤智義* 干川尚人†
Harutaro Abe Kaito Watanabe Hirotaka Nakayama Atsushi Shiraki Tomoyoshi Ito Naoto Hoshikawa

1. はじめに

近年のメタバースやゲーム，シミュレーション分野の発展に伴い，3DCG（3- Dimensional Computer Graphics）コンテンツは様々な領域で利用されており，その結果， 3D モデルの需要は年々高まっている．しかし，一般的に 3D モデルを作成する際には，Unity[1]や blender[2]などの 3D モデリングソフトが用いられるが，これらの扱いは非常に複雑である．そのため，使用者には専門的な知識と経験が必要であり，このことから深刻なクリエイター不足が懸念されている．加えて，熟練者であっても，一つの造形物を作成するためには，パーツとなる小さなオブジェクトを一つ一つ配置してから変形，着色などの加工が必要になるため，非常に時間と費用がかかる．また，3D モデルの中でもキャラクターやアイテムといった小型のオブジェクトはインターネット上で多数配布されており， AI （Artificial intelligence）による自動生成技術も発達している．しかし，山や海岸といった大型の背景オブジェクトは，インターネット上での配布は少なく，自動生成技術もあまり発達していない．加えて小型オブジェクトに比べて大規模であるため，制作に多大なコストが必要である一方で，開発者やデザイナーから重要視されづらく，コストを十分に割けないという課題がある．

これらの課題に対し，3D フラクタルモデルを利用し自然背景モデルに特化した 3D モデルの自動生成システムが提案されている [3][4]．しかし，そのシステムでは入力として，自然背景の写真やイラストなどの 2D（2-Dimensional）データを必要とするため，ユーザーの意図を柔軟に反映できない課題があった．この課題に対し本研究では，自然言語による入力を用いた柔軟な操作が可能な 3D 背景モデルの自動生成システムを提案する．

2. 関連技術および関連研究

2.1 コンテンツ生成 AI

近年，ユーザーの入力に基づき 3D モデルやイラストなど，様々な種類のコンテンツを生成する AI の普及が進んでいる．画像生成においては，Adobe の「Adobe Firefly」[5]，NVIDIA の「GauGAN」[6]，OpenAI の「DALL・E3」[7]などがあり，これらはユーザーから受け取ったテキストやスケッチなどの情報から，新しい画像を自動生成することが可能である．また，3D モデルを生成する技術については，Google Research とカリフォルニア大学バークレー校が開発した「Zero-Shot Text-Guided Object Generation with Dream Fields」[8]などが挙げられる．これはテキストから 3D モデルを生成できるが，目標物は花瓶や小型生物など，比較的小さなオブジェクトの生成に限定される．加えて，精度の高い 3D モデルを生成するためには，2D の画像生成技術と比べ膨大な量の学習データが必要となる．したがって，高品質かつ巨大な 3D モデルを生成する技術は充分に確立していないという問題がある．

2.2 3D フラクタルモデル生成システム

巨大な 3D モデルを生成する技術として，フラクタルモデルを用いた 3D 自然背景モデルの自動生成手法が提案されている[3][4]．フラクタル（Fractal）とは，自己相似性と呼ばれる，細部を拡大したものと全体像が図形的に類似するという幾何学的概念であり，代表的なものとしてはマンデルブロ集合やメンガーのスポンジが挙げられる．また，山の稜線，海岸線など自然背景の多くがフラクタル図形で近似することができると知られている[9]．任意の 3D フラクタルモデルを生成するためには，多数の生成用パラメータをフラクタル生成エンジン[10]に入力する必要がある．この研究では，そのパラメータを遺伝的アルゴリズム（Genetic Algorithm: GA）[11]によって最適化することで，ユーザーの意図に沿った 3D モデルを生成することを目指している．この遺伝的アルゴリズムの適応度には，生成された 3D フラクタルモデルの形状とユーザーが入力した画像との類似度を利用していた．しかし，この適応度評価手法ではユーザーのイメージ情報として，自然背景の写真やイラストなどを入力する必要があるため，ユーザーの意図を正確に反映することが困難であった．例えば，「険しい山とその麓に湖」や「画面右半分にまばらに木を配置する」など，具体的な指示は画像だけでは伝えることが困難である．結果として，ユーザーのイメージを柔軟にシステムに入力することは難しく，生成される 3D モデルに利用者の意図が正確に反映されないという問題があった．

3. 提案手法

既存の 3D フラクタルモデル生成システム[3][4]では，ユーザーのイメージを柔軟に入力することは難しく，ユーザーの意図を正確に反映した 3D オブジェクトの生成が困難であるという課題があった．この原因として，既存の適応度評価手法では，入力画像とレンダリング画像の類似度を利用していたため，システムの入力手段が画像に限られていたことが挙げられる．この課題を解決するため，本研究では，自然言語を用いた柔軟な入力によるテキスト駆動型の 3D フラクタルモデルの自動生成手法を提案する．既存システムでは，事前学習済みモデルの VGG-19[12]を用いた適応度評価手法が利用されていた．この手法では，入力画像とレンダリング画像を学習済みモデルのフィルターに通すことで特徴量ベクトルを抽出し，このベクトル間の距離を求めることで，画像間の類似度を算出していた．提案手法では，ユーザーが入力したテキストと生成画像の類似度を適応度として利用する．この評価手法では，図 1 に示すように，ユーザーが入力したテキストとシステムによる生

† 筑波大学 University of Tsukuba
‡ 国立天文台 National Astronomical Observatory of Japan
＊千葉大学 Chiba University

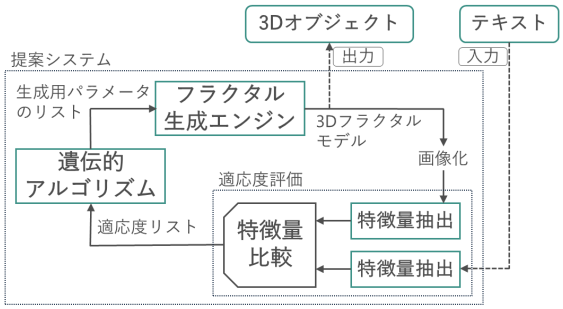

図 1　提案手法の概要

成画像を学習済みモデルに入力し，特徴量ベクトルを抽出する．その後，ベクトル間のコサイン類似度を求めることで，テキストと画像間の類似性を評価でき，その値を適応度として利用する．この評価手法を基に，遺伝的アルゴリズムを用いて生成用パラメータを最適化することで，テキスト入力から 3D フラクタルモデルの自動生成が可能となる．本研究では，この適応度評価を「テキスト駆動型適応度評価手法（Text-Driven Fitness Evaluation：TDFE）」と呼称する．具体的には，マルチモーダルな学習済みモデルである，OpenAI の CLIP[13]を用いることで，テキストと画像間の特徴量抽出を可能にし，TDFE を実現している．TDFE を既存システムに追加することで，ユーザーは意図的なテキスト指示に基づく 3D フラクタルモデルの生成を試みることができる．例えば，既存システムでは「中央が大きく割れた山」と類似する 3D フラクタルモデルを生成するためには，ユーザーはそのイメージを表すイラストや写真などの画像を事前に準備する必要があった．そのため，このような複雑な風景に完全に合致するイメージ画像を用意するには長い時間を要し，入手自体が困難であった．これに対し，本システムでは，「中央が大きく割れた山」というテキストをそのままシステムに入力することができるため，ユーザーには柔軟な入力が可能となる．

4. システム実装

本章では，提案する TDFE の有効性を確認する機能評価を行う．まず，4.1 節では CLIP の文字入力による類似度計算結果が 3D フラクタルモデル生成の適応度評価機能に組み込めるかを検証する．4.2 節では，これが意図通り 3D フラクタルモデルを出力できるかを検証する．

4.1 CLIP による TDFE の類似度出力

類似度計算の検証では，テキストと画像の組を入力し，出力される類似度を確認する．ここでは，入力するテキストとして単語と句それぞれで実行する． この検証で TDFE に画像では表現しきれないような詳細な指示を文字情報で与えても，正しく類似度評価が行えるかを確認する．

まず，名詞の単語と画像間の類似度評価を検証する．検証に用いたデータは，図 2 に示す(a)MOUNTAIN，(b)CAT，(c)BUILDING の 3 カテゴリに属する画像 5 枚ずつと，各カテゴリに対応する名詞（"mountain"，"cat"，"building"）である．実験の結果を表 1 に示す．表 1 の各行は各カテゴリに属する画像 5 枚の平均類似度を表している．表中の類似度は既存システム[3][4]に倣って，値が小さいほど類似したデータ同士であることを表している．得られた結果より，

図 2　検証に用いた画像データ

表 1　テキストと画像間の類似度

	mountain	cat	building
(a)MOUNTAIN	**3.75**	5.68	4.85
(b)CAT	4.97	**3.65**	5.31
(c)BUILDING	5.10	6.10	**4.26**

画像とテキストの組が合致しているとき得られた類似度は，合致していない場合に比べて小さいことが確認できた．

次に，名詞の単語に"Snow-covered", "Steep", "towering"などの形容詞または形容詞句を修飾語とした名詞句を作成し，画像と文章間の類似度を算出し，その値がどのように変化するのかを確認する．検証に用いたデータは，図 2 の(a)MOUNTAIN に属する画像 5 枚と，それぞれの画像に合致した，画像の内容を説明する文章である．与えた文章の例として，図 2(a)MOUNTAIN の左端の画像に合致する文章は"Snow-covered mountain."である．これらの文章は，主観を排除するため，OpenAI の ChatGPT 4o[14]に画像を入力し，その構造的特徴について説明するように指示を与え，生成された文章を使用している．この文書によって得られた類似度と，単語である"mountain"によって得られた類似度を比較する．この結果，文章と画像を入力した場合，平均類似度は 3.67 となった．また，単語を入力した場合の平均類似度は，表 1 より 3.75 である．つまり，単語を入力したときと比べ，文章の方が，類似度の値が小さくなっている．この結果によって，一つの単語だけでなく，複数の何語から成る句を入力した場合でも，類似度評価が実行できることを確認できた．また，名詞を形容詞や形容詞句で追加の情報付与することで，評価の正確性が上がる可能性も示唆している．

以上の結果はテキストと画像の類似性を判定する CLIP の機能が意図通り動くことを示す．これにより生成される画像の評価において，目的とする画像の内容をユーザーが文字列で指示可能になり，TDFE の基本機能を確認できたと言える．

4.2 TDFE の実装評価

本節では，TDFE を実装した 3D フラクタルモデル生成システムで，生成用パラメータを最適化し，実際に 3D フラ

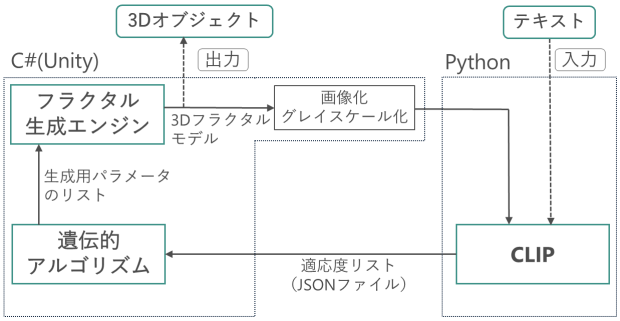

図 3　実験システム

表 2　RCGA の設定値

Setting Item	Setting Value
Max Generation	1000
Num Parameters	6
Init Param Range	+4.0 ~ -4.0
Population	150
Parent	7
Child	90
Elite	3
Mutation Rate(%)	5

クタルモデルを生成することで，システムが正しく動作することを評価する．

4.2.1 実験システムおよび実験条件

TDFE を用いて実験システムを構築する．3D フラクタルモデル生成システムは，フラクタル生成エンジン，遺伝的アルゴリズム，適応度評価の 3 つのモジュールから構成されている．実験システムでは，既存システムの適応度評価モジュールに TDFE に変更することで，ユーザーのテキスト入力から 3D フラクタルモデルを自動生成することを可能にしている．TDFE は生成画像とテキストを CLIP に入力することで，特徴量ベクトルを抽出し，これらのベクトルを比較することで，類似度を算出している．ただし，この類似度は，既存システムに倣って，値が小さいほど，入力データ同士がより類似していることを示す．また，CLIP は python のライブラリで公開されているため，TDFE は python によって実装する．次に，既存システムにおけるフラクタル生成エンジンは Unity を基に実装されている．また，遺伝的アルゴリズムは Unity に組み込む形で実装され，木村らによる実数値データ向けの遺伝的アルゴリズムである，実数値遺伝的アルゴリズム (RCGA：Real-Coded GA) [15] を利用していた．図 3 に実験システムにおける各モジュールの接続について示す．図 3 に示すように，これらのモジュールは以下の手順に従って接続されている．

1) フラクタル生成エンジンで生成用パラメータから 3D フラクタルモデルを生成し，画像出力する．
2) 出力された画像は，3D 形状にのみ着目するため，グレイスケール化を行う．
3) これらの画像を TDFE で受け取り，事前入力されたテキストとの類似度を算出する．
4) この類似度から適応度リストを作成し，json 形式のファイルとして，遺伝的アルゴリズムに渡す．
5) 遺伝的アルゴリズムでは，適応度リストを用いて生成用パラメータの最適化を行う．

以上の処理を繰り返し行うことで，生成用パラメータの最適化が進行し，3D フラクタルモデルが生成される．

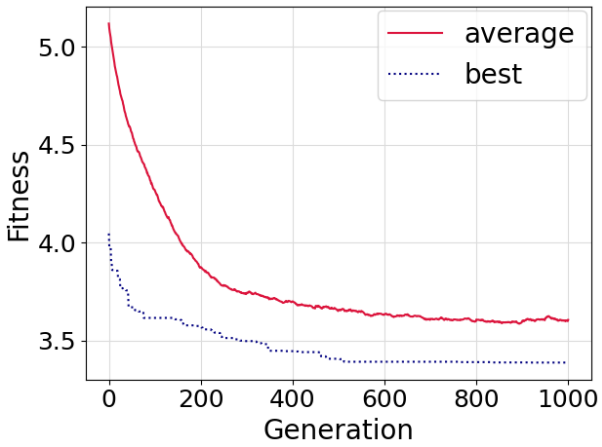

図 4　各世代における適応度の推移

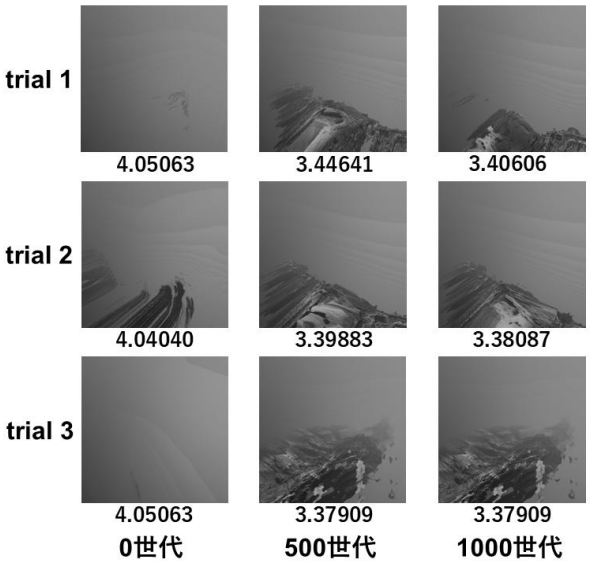

図 5　最適化過程における生成画像

次に，実験に使用した実数値遺伝的アルゴリズムの設定値を表 2 に示す．これらの値は，先行研究[3][8]と同様に，木村らによる解説[15]の推奨値を参考にしている．また，フラクタル生成エンジンに含まれる，形状決定に関わる生成用パラメータは全 15 種であるが，今回の実験では 6 つを最適化の対象にする．これは先行研究[3][4]において，高品質な最適化が可能であることが明らかとなっていることが理由である．今回の実験は，以上の条件で 3 回試行を行なっている．

次に，入力するテキストは，単語に形容詞句など修飾語を付与して作成された文を用いる．実験に使用した文章について，以下に示す．

- "On the lower side of the image is a large steep mountain, the slopes of which are bumpy and rocky. A lake has formed in the lower right.".

4.2.2 実験結果および考察.

実験で得られた結果は図 4 と図 5 の通りである．図 4 は遺伝的アルゴリズムによる適応度値の変化を示すグラフであり，3 回の試行の平均を表している．横軸は世代数，縦軸は適応度を表している．また，凡例の average は遺伝的

アルゴリズムの母集団に含まれる全個体の平均適応度，best は母集団中の最も適応している個体の適応度を示している．また図 5 の行は各試行を，列は遺伝的アルゴリズムの世代を表しており，画像はその世代における最も適応している個体，下の数値はその適応度を表している．

図 4 に示す適応度の推移を確認すると，エリート保存により最適な個体の適応度は着実に減少し続けており，母集団中の全個体の平均適応度も安定して収束している．また，700 世代あたりから平均適応度の値はほぼ変化しなくなっていることから，このあたりで値はおおむね収束していることがわかる．このことから，生成用パラメータの最適化は問題なく進んでいることに加え，遺伝的アルゴリズムの世代数は充分であることが確認できた． 次に，図 5 の各試行で生成された画像を確認すると，全ての試行において，世代を重ねるごとに徐々にオブジェクトが形成されていく様子が確認できる．また，最終世代（1000 世代目）において，"山"に類似したオブジェクトが確認できる．例えば，図 5 の trial1 を確認すると，第 0 世代でははっきりとした構造は何も確認できない．しかし第 500，1000 世代と進むにつれ，画面下部に山の構造を持つオブジェクトが形成された．この結果から，テキスト入力から TDFE を用いて 3D フラクタルモデルを自動生成できることが示された．

以上より，TDFE を組み込んで生成された 3D フラクタルモデルは，実用的なオブジェクトの生成にはまだまだ課題があるが，既存研究の VGG-19 による処理と同程度の品質での出力は可能であることを確認できた．

5. 今後の課題

前章までの評価結果により，提案システムは 3D フラクタルモデルの生成にテキスト入力による指示が可能であることを示したが，一方で本システムが生成するオブジェクトがどの程度ユーザーの意図を的確に反映しているか，その検証が課題である．これを評価するためには，本研究で実施した特定の目的，単語，句や文だけでなく，より幅広い目的やバリエーションの豊富な表現を組み合わせて，ユーザーの意図に沿う出力に繋がっているかを確認する必要がある．

また，今回の実験結果は実用観点での品質には達していないため，その改善のためにはより複雑なフラクタル構造の生成が必要で，これには生成パラメータの数を増やす必要がある．しかし，安易なパラメータ拡張は組合せ爆発を引き起こし，収束が困難になるため，その解決方法が課題である．

6. おわりに

近年，3D 背景モデルの需要は年々高まっているが，その制作には高いコストがかかるという課題がある．本研究では，この課題に対して，テキストによる指示に基づいて背景モデルを自動生成するシステムを提案した．従来のシステムでは，入力として自然背景の写真などを必要としていたため，ユーザーの意図を柔軟に反映することが困難であった．そこで本研究では，テキストと画像間の適応度評価手法である TDFE を用いて，テキストによる柔軟な操作を可能にした．さらに，提案手法を既存の 3D フラクタルモデル生成システムに組み込み，実験システムを構築してその動作を検証した．その結果，システムは問題なく動作することが確認された．一方で，本システムの性能に関する評価は未実施であるため，どれほどユーザーの意図に沿ったオブジェクトが生成できているか示されていないという課題が存在する．このため，様々なテキストを入力しての実験や，生成用パラメータ数の増加などが必要である．

謝辞

本研究は JSPS 科研費 JP23K11078 ならびに，筑波大学研究基盤支援プログラム（S タイプ）（University of Tsukuba Basic Research Support Program Type S）の支援を受けて行われている．

参考文献

[1] Unity Technologies, “Unity Real-Time Development Platform”, Unity Real-Time Development Platform, Available: https://unity.com/, (Accessed: May.2025)
[2] Blender Foundation., “blender”, blender, Available: https://www.blender.org/, (Accessed: May.2025)
[3] K. Watanabe, H. Nakayama, A. Shiraki, T. Ito, K. Hirata, and N. Hoshikawa, “A Method for Automatic Generation of 3D Background Models Using Fractal Models and Genetic Algorithms”, 2023 Eleventh International Symposium on Computing and Networking Workshops (CANDARW), pp.135-141, Nov.2023, DOI: 10.1109/CANDARW60564.2023.00030.
[4] 渡邊海斗, 千川尚人, 中山弘敬, 伊藤智義, 白木厚司, “フラクタルモデルと遺伝的アルゴリズムによる 3D 背景モデル自動生成手法,” 信学技報, vol. 122, no. 406, NS2022-190, pp. 133-138, 2023
[5] Adobe, “Adobe Firefly”, Adobe Firefly, Available: https://www.adobe.com/jp/sensei/generative-ai.html, (Accessed: May.2025)
[6] NVIDIA Corporation, “‘Paint Me a Picture’: NVIDIA Research Shows GauGAN AI Art Demo Now Responds to Words”, NVIDIA Blog, Available: https://blogs.nvidia.com/blog/gaugan2-ai-art-demo/, (Accessed: May.2025)
[7] OpenAI,Inc., “Research DALL·E 3”, DALL·E 3, Available: https://openai.com/index/dall-e-3/, (Accessed: May.2025)
[8] A. Jain1, B. Mildenhall, J. T. Barron, P. Abbeel, and B Poole, “Zero-Shot Text-Guided Object Generation with Dream Fields”, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.857-866, Jun.2022, DOI: 10.1109/CVPR52688.2022.00094.
[9] B.B.Mandelbrot, Fractals: form, chance, and dimension. New York, NY, USA: W.H.Freeman and Company, 1977.
[10] H. Nakayama, and SIZIMA Soft, “Beanstalk version 1.0 b2”, Beanstalk, Available: https://sizima.com/beanstalk/beanstalk.html, (Accessed: May.2025)
[11] D. E. Goldberg, Genetic Algorithms in search, Optimization, and Machine Learning. Hoboken, NJ, USA: Addison-Wesley, 1989.
[12] K. Simonyan, and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, International Conference on Learning Representations (ICLR), Sep.2014, DOI: 10.48550/arXiv.1409.1556
[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision”, arXiv, DOI: 10.48550/arXiv.2103.00020
[14] OpenAI, Inc., “GPT-4o が登場”, GPT-4o が登場, Available: https://openai.com/ja-JP/index/hello-gpt-4o/, (Accessed: May.2025)
[15] 木村元, “連続な多次元変数の最適化：実数値遺伝的アルゴリズム 世代交代モデル JGG ＋ 多親交叉 REX”, 連続な多次元変数の最適化：実数値遺伝的アルゴリズム 世代交代モデル JGG ＋ 多親交叉 REX, Available: http://sysplan.nams.kyushu-u.ac.jp/gen/edu/Algorithms/RGA_JGGandREX/RGA_JGG_REX.html, (Accessed: May.2025)