# 任意のメッセージ長に対応した生成 AI 画像向け電子透かし技術の開発

Development of a Watermarking Technology for Generative AI Images with Arbitrary Message Length



**川村　正樹**（Masaki KAWAMURA, Ph. D.）
**山口大学　大学院　創成科学研究科　教授**
(Professor, Graduate School of Sciences and Technology for Innovation, Yamaguchi University)
電子情報通信学会（フェロー）　IEEE　APSIPA　他
受賞：電子情報通信学会 ISS 功労賞 (2017 年)
著書：CentOS8 で作るネットワークサーバ構築ガイド, 秀和システム, 2020 年　他
研究専門分野：ニューラルネットワーク　電子透かし

**あらまし**

画像生成 AI の普及に伴い、人の目では真偽を判別できない精巧な偽画像の拡散が社会問題となっており、生成画像の出所を証明する電子透かし技術の重要性が増している。従来の後処理型の透かしは、画像を拡散モデルで再構成するだけで容易に消去されてしまうため、生成過程に透かしを組み込む手法 DiffuseTrace が提案された。しかし、この手法は、埋め込むメッセージの長さを変更するたびに専用の VAE を再学習する必要があり、実運用上の負担が大きかった。本研究では、学習を必要としない一層構造のランダムネットワークを用いてメッセージを透かしパターンへ変換する手法 FlexDT を提案した。これにより、VAE を再学習することなく任意の長さのメッセージに対応可能となった。実験の結果、抽出精度は Acc=0.993 となり、従来手法（Acc=0.965）を上回るとともに、画像品質にも有意な劣化が見られないことを確認した。

## １．研究の目的

画像や映像、音楽などのディジタルコンテンツを保護する技術として、電子透かし法がある。この手法の特徴は、コンテンツ自体はそのまま利用でき、その品質をほとんど低下させずに、著作権情報や購入者 ID などの透かし情報をこっそりと埋め込ませることである。生成 AI が登場する以前の画像に対する典型的な電子透かしの利用法は、つぎのとおりである。まず、価値のある画像などが先にあり、事後的に透かし情報を埋め込む処理が行われる。不正利用者は画像に対してノイズの付加、圧縮、幾何変換、切抜きなどを行い、自身の購入者 ID を消そうと試みる。著作権者は加工された画像から透かし情報を抽出し、不正利用者を特定する。このような高耐性電子透かし法が必要とされていた。

生成 AI が登場すると、この状況は一変する。後処理型の透かし埋め込み法では、画像に一度ノイズを加えてから拡散モデルで再構成するだけで容易に消去できてしまうことが指摘されており、堅牢性に大きな課題を抱えている。そのため、生成 AI に対応した電子透かし法の開発が必要となった。これより、本研究の目的は、生成 AI で作成した画像に透かし情報を埋め込む方法と抽出する方法を開発することである。

## ２．研究の背景

生成 AI はプロンプトに従い、高精細な画像を生成できる。既に人が見ただけでは自然画像と生成画像を判別することは困難になっており、フェイク画像の拡散が社会問題になっている。そのため、画像の真正性を示すためにも画像に透かし情報を埋め込むことの重要性が再認識されるようになった。生成 AI 内部への透かし埋め込みは、既に実用化が始まっている。例えば Google は、自社の画像生成 AI が出力する画像に対して、目に見えない電子透かし技術 SynthID [1] を組み込んでおり、ユーザーは Gemini アプリ上で画像が Google の AI によって生成されたものかどうかを検証できる。

画像生成 AI の中核をなす技術に、潜在拡散モデル (Latent Diffusion Model：LDM) がある [2]。LDM は、画像を一旦 VAE（変分オートエンコーダ）[3] と呼ばれるネットワークで圧縮し、その圧縮された表現（潜在空間）上でノイズの除去を繰り返すことで画像を生成する仕組みであり、Stable Diffusion や DALL-E などの代表的な画像生成 AI に採用されている。ピクセル

単位で処理する従来の拡散モデルに比べ、計算コストを大幅に削減できる点が特長である。

この LDM の生成プロセスの内部に透かしを組み込む手法として、DiffuseTrace [4] が提案されている。DiffuseTrace は、埋め込みたいメッセージ（ビット列）を画像生成の出発点となる初期ノイズに変換し、そのノイズから画像を生成する。生成された画像からは、逆拡散処理*1 と呼ばれる初期ノイズを復元する処理によって、埋め込んだメッセージを取り出すことができる。この方式は、透かし専用の VAE（以降、透かし用 VAE と呼ぶ）を介してメッセージと初期ノイズを相互に変換する点に特徴がある。なお、前述の SynthID が生成画像のピクセル値に直接変化を加える方式であるのに対し、DiffuseTrace は生成プロセスの初期ノイズに透かしを埋め込む方式であり、両者はアプローチが異なる。

我々は任意のメッセージを埋め込める DiffuseTrace を取り上げ、メッセージを変換する処理に着目した。このモデルでは、メッセージの各ビットを $64 \times 64$ 個のビット列に変換していた。例えば、例えば、ビット '0' を $64 \times 64$ 個の'0' に変換していた。また、DiffuseTrace [4] には重大な制約があった。透かし用 VAE の入力の次元（チャンネル数）がメッセージの長さと一致するように設計されているため、埋め込みたいメッセージの長さを変更するたびに、透かし用 VAE 全体を作り直し、再学習する必要があったのである。これは実運用上、大きな時間的・計算資源的コストとなる。我々はもっと効率的な符号化ができるのではないか、また、メッセージ長を変えても再学習が不要にできないかを検討した。

## ３．提案モデル

本研究では、メッセージの長さに依存しない固定サイズの透かしパターンへの変換を、学習を必要としないランダムネットワークを用いた一層構造の符号化層によって実現する手法 FlexDT を提案した [5]。図 1 に FlexDT の符号化器と復号器を示す。

### 3.1 ネットワーク構造

メッセージを透かし情報に変換し、透かし入り画像を生成する埋め込み過程と、画像からメッセージを抽出する過程の概略を説明する。埋め込み過程では、$N$ ビットのメッセージ $\boldsymbol{m}$ を提案するランダムネットワークで変換し、透かし情報とする。これを透かし用 VAE に入力する。透かし用 VAE の入力層は $64 \times 64 \times C$ 次元であり、$C$ はチャネル数である。透かし用 VAE は、DiffuseTrace で用いられていた事前学習済みの VAE である。VAE の出力が Stable Diffusion の潜在空間の初期値となり、画像が生成される。透かし情報の抽出過程では、出力側から画像が入力され、逆拡散処理によって、潜在空間に射影される。この値を透かし用 VAE の復号器にかけて、透かし情報を得る。提案するランダムネットワークを用いて、メッセージが推定される。

### 3.2 メッセージの符号化器

FlexDT では $N$ ビットのメッセージ $\boldsymbol{m}$ は符号化用のランダムネットワークによって、$64 \times 64$ 次元のベクトル $\boldsymbol{\phi}$ に符号化される。つまり、$\boldsymbol{\phi} = Q\boldsymbol{m}$ である。ここで、提案ネットワークは一層の構造であり、結合重み行列 $Q$ は列直交行列*2 である。この行列 $Q$ は任意の値を取る行列 W を QR 分解することで得られ、固定値を取るため、学習する必要はない。透かし用 VAE に入力するために、チャネル数回ベクトル $\boldsymbol{\phi}$ を複製し、$64 \times 64 \times C$ 次元の透かし情報 M を得る。この様に、メッセージ長 $N$ とチャネル数 $C$ は独立に選べるようにしたので、メッセージ長が変わっても透かし用 VAE の再学習は不要にできた。

### 3.3 メッセージの復号器

生成画像から逆拡散処理によって得られた潜在変数は、$C$ チャネル分ある。そのため、チャネル間で平均をとり、$64 \times 64$ 次元のベクトル $\overline{\boldsymbol{\phi}}$ を得る。これを復号用のランダムネットワークに入力する。このネットワークの結合重み行列は $Q^\top$ であり、直交行列の性質により、$Q^\top Q = I$ を満たす。すなわち、$\widehat{\boldsymbol{m}} = Q^\top \overline{\boldsymbol{\phi}}$ であり、$\widehat{\boldsymbol{m}} = Q^\top Q\boldsymbol{m} = \boldsymbol{m}$ より、画像に劣化が無ければメッセージを復元可能である。

以上のように、 FlexDT はランダムネットワークの性質を利用し、可変長のメッセージに対応し、ランダムネットワークと透かし用 VAE の再学習が不要である特長を持っている。

# 任意のメッセージ長に対応した生成 AI 画像向け電子透かし技術の開発

Development of a Watermarking Technology for Generative AI Images with Arbitrary Message Length

## 4．モデルの評価と結果

提案手法である FlexDT の性能を評価する。透かし用 VAE と LDM は、事前学習済みの DiffuseTrace を利用し、メッセージの符号化器と復号器のみを実装した。画像生成用の LDM として Stable Diffusion v2-1-base を用い、約 8 万件のプロンプトからなる Diffusion Prompts データセットから抜き出した 50 件のプロンプトを用いて画像を生成した。計算機環境には、本課題において購入した CPU: Intel Core i9、GPU: NVIDIA GeForce RTX 4090、メモリ 32GB を搭載した計算機を用いた。透かし抽出精度は、元のメッセージと抽出したメッセージの一致度である精度 (Acc)によって評価する。また、生成画像の画質は、NIQE*3 と PIQE*4 によって評価する。これらの指標は画像の自然さや知覚的な品質を数値化した指標である。生成画像には原画像が存在しないため、PSNR や SSIM では評価できないことに注意する。さらに、生成画像とプロンプトの意味的な一致度を測る CLIP スコア*5 も用いた。

### 4.1 性能比較

DiffuseTrace と精度と画質を比較するために、メッセージ長とチャネル数をどちらも 48 に設定した。このときの FlexDT のメッセージの精度は Acc=0.993 であり、DiffuseTrace の Acc=0.965 よりも上回る結果を得た。DiffuseTrace は、メッセージの各ビットをそれぞれ独立したチャンネルに割り当て、$64 \times 64$個の同一値に複製することで冗長化している。これに対し FlexDT は、$N$ ビットのメッセージ全体を列直交行列によって$64 \times 64$の 1 枚のパターンへ一括して変換したうえで、これを $C$ チャンネルに複製している。復号時には、チャンネルで平均してノイズを抑えた後、列直交行列の転置によって元のメッセージへ逆変換するため、チャンネル方向の平均化と行列演算による重ね合わせの両方でノイズの影響が打ち消され、これが精度向上につながったと考えられる。

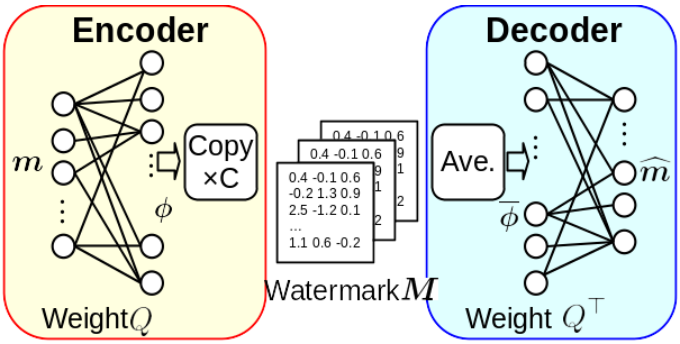



図 1 FlexDT の符号化器(Encoder)と復号器(Decoder)



図 2 (左)透かし情報の埋め込みなしの生成画像、(右)提案手法で透かし情報を埋め込んだ生成画像 [5]

図 2 に生成した画像を示す。同じプロンプトでも初期値の違いにより、異なる画像が生成される。画像品質については、NIQE・PIQE・CLIP スコアのいずれにおいても、透かしを埋め込まない画像と比べて有意な劣化は見られなかった。これは透かし由来の情報を潜在空間の初期値として与えているため、その後生成される画像の画質に影響しないからである。以上より、透かし情報の埋め込みが画像の見た目や意味内容に悪影響を及ぼさないことが確認された。

### 4.2 様々なメッセージ長に対する精度

FlexDT はメッセージ長を変更しても、VAE の再学習は不要である。そこで、同一の VAE を用いて、様々なメッセージ長に対する精度を評価した。チャネル数は $C = 48$ で固定とする。。図 3 はメッセージ長に対する精度を表している。なお、比較のため二層構造の FlexDT [6] も構築し、その結果を併記した。その結果、1024 ビット以下であれば、Acc=0.965 よりも上回る精度を維持できることが分かった。また、2048 ビットでも Acc>0.9 であり、十分に長いメッセ

ージを埋め込むことができることが分かった。

### 4.3 様々な攻撃に対する性能評価

透かし情報を埋め込んだ生成画像に対して、輝度値の変更、ガウスノイズの付加、コントラスト変調、JPEG 圧縮、リサイズなどの画像加工を加えた後、抽出したメッセージの精度を評価した。FlexDT は全体として DiffuseTrace と同等以上の精度を示すことが確認された。FlexDT は全体として DiffuseTrace と同等以上の精度を示すことが確認された。特に、$N = 1024$ ビットのような比較的長いメッセージにおいても、この頑健性が維持されることを確認した。

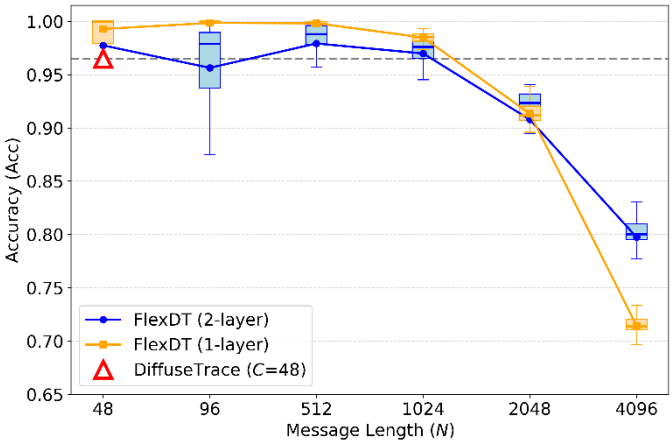



図 3　様々なメッセージ長に対する精度 [5]

## 5. 将来展望

本研究により、画像生成 AI に組み込む電子透かしについて、メッセージ長を変更するたびに専用ネットワークを再学習する必要がなく、かつ従来より少ない計算量で高い抽出精度を実現できることが示された。これは、画像生成サービスの提供事業者が、利用者ごとに異なる識別コードをリアルタイムに、かつ低コストで埋め込むといった実運用の場面で特に有用と考えられる。

今後は、Stable Diffusion 以外の様々な画像生成 AI や、動画生成 AI など画像以外のモダリティへの適用可能性についても検討していく予定である。これらを通じて、生成 AI の出所を確実に追跡できる仕組みを社会に広く普及させたいと考えている。

## おわりに

本稿では、生成 AI による画像に対して、再学習を必要とせずに任意の長さのメッセージを埋め込み・抽出できる電子透かし手法 FlexDT を紹介した。一層構造のシンプルなランダムネットワークを用いるだけで、従来手法を上回る精度と、画像品質を損なわない実用性を両立できることを示せたと考えている。今後は、より長いメッセージへの対応や、様々な生成 AI サービスへの実装を通じて、生成 AI の信頼性向上に貢献していきたい。

## 謝辞

本研究は、SCAT（テレコム先端技術研究支援センター）の研究費助成をはじめ、JSPS 科研費（課題番号 JP20K11973、JP24K15106）による支援を受けて実施したものである。

## 用語解説

*1 逆拡散処理：拡散モデルが画像を生成する際に、ノイズから徐々に画像を復元していく処理である。この処理を逆にたどることで、生成された画像から生成の出発点であった初期ノイズを復元することもできる。

＊2 列直交行列：転置した行列を掛け合わせると単位行列になるという性質を持つ行列である。この性質により、画像に劣化がなければ、埋め込んだメッセージを誤差なく復元できることが理論的に保証される。

＊3 NIQE (Natural Image Quality Evaluator) ：あらかじめ学習しておいた自然画像の統計的な特徴との違いから、参照画像なしで画像の自然さを評価する指標である。値が小さいほど自然な画像であることを示す。

＊4 PIQE (Perceptual Image Quality Evaluator)： 画像を小さなブロックに分割し、ブロックごとの局所的な劣化の特徴から知覚的な画質を評価する指標である。値が小さいほど高品質であることを示す。

＊5 CLIP スコア：画像とテキストを同じ意味空間に対応付けるよう学習された画像・言語モデル CLIP を用いて、生成画像とプロンプトの意味的な一致

度を数値化した指標である。値が大きいほど一致度が高いことを示す。

**参考文献**

[1] S. Gowal, R. Bunel, F. Stimberg, D. Stutz, et al., "SynthID-Image: Image watermarking at internet scale," Google DeepMind, arXiv:2510.09263, 2025.
[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10684–10695, 2022.
[3] D.P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv:1312.6114, 2013.
[4] L. Lei, K. Gai, J. Yu, and L. Zhu, "DiffuseTrace: A transparent and flexible watermarking scheme for latent diffusion model," arXiv:2405.02696, 2024.
[5] O. Egami and M. Kawamura, "FlexDT: Flexible and Retraining-free Variable-length Encoder-Decoder for DiffuseTrace," IEICE Transactions on Information and Systems, E110-D (4), 2027.（掲載予定）.
[6] O. Egami and M. Kawamura, "Proposal of a random encoding layer compatible with arbitrary message lengths for diffusetrace," 2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.2170–2175, IEEE, 2025.

**関連文献**

[1] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-Rings watermarks: Invisible fingerprints for diffusion images," Advances in Neural Information Processing Systems, vol.36, pp.58047–58063, 2023.
[2] X. Zhao, K. Zhang, Z. Su, S. Vasan, I. Grishchenko, C. Kruegel, G. Vigna, Y.X. Wang, and L. Li, "Invisible image watermarks are provably removable using generative AI," Advances in Neural Information Processing Systems, vol.37, pp.8643–8672, 2024.
[3] 江上旺, 川村正樹, "DiffuseTrace のための任意のメッセージ長に対応した符号化層の設計と評価," 電子情報通信学会第 2 回 EMM 研究会, 2025 年 7 月.
[4] 江上旺, 川村正樹, "DiffuseTrace のための単層可変長符号化ネットワークと性能評価," 電子情報通信学会 EMM 研究会, 2026 年 3 月.

この研究は、令和５年度ＳＣＡＴ研究助成の対象として採用され、令和６～７年度に実施されたものです。