Blue Noise for Diffusion Models

Xingchang Huang　MPI Informatics, VIA Center, Germany
Corentin Salaün
MPI Informatics, Germany　Google DeepMind, UK
Cristina Vasconcelos　Christian Theobalt
MPI Informatics, VIA Center, Germany
Cengiz Öztireli
Google Research, University of Cambridge, UK
Gurprit Singh
MPI Informatics, VIA Center, Germany

概要

現在の拡散モデルの多くは、すべての時間ステップにおいてガウスノイズを使用してトレーニングおよびサンプリングを行っています。しかし、この方法では、デノイジングネットワークによって再構築される周波数成分を最適に考慮しているとは限りません。コンピュータグラフィックスにおける相関ノイズの多様な応用にもかかわらず、そのトレーニングプロセス向上への可能性は十分に探求されていません。

本研究では、画像内および画像間で相関ノイズを考慮する新しい一般的なクラスの拡散モデルを提案します。具体的には、トレーニングプロセスに相関ノイズを組み込むための時間変化ノイズモデルと、高速に相関ノイズマスクを生成する方法を導入します。

本モデルは、決定論的拡散モデルに基づいており、ガウス白色（ランダム）ノイズのみに比べてブルーノイズを活用することで生成品質を向上させます。さらに、本フレームワークは単一のミニバッチ内で画像間の相関を導入することを可能にし、勾配フローを改善します。

本手法を用いて、さまざまなデータセットで定性的および定量的評価を行い、既存の決定論的拡散モデルを FID（Fréchet Inception Distance）メトリックで上回る結果を達成しました。コードは以下のリンクで公開される予定です：https://github.com/xchhuang/bndm。

1　はじめに

Sohl-Dickstein ら [2015]、Ho ら [2020]、および Song と Ermon [2019] の画期的な研究以来、拡散モデルに関する研究が盛んに行われています。これらのモデルは、生成品質およびトレーニングの安定性の観点で、Generative Adversarial Networks（GANs）[Dhariwal and Nichol 2021] を上回る性能を示しており、画像生成において優れた成果を挙げています。また、拡散モデルは、テキストから画像の生成、画像のインペインティング、画像の超解像、画像編集といった様々なタスクを実行するように訓練することも可能です。

通常、拡散モデルは**順方向プロセス**と**逆方向プロセス**の 2 つのプロセスで構成されます。順方向プロセスでは、モデルが元のデータポイント（例：画像）にノイズを徐々に加え、それをランダムなノイズパターンに変換します。一方、逆方向プロセスでは、デノイジングニューラルネットワークを用いて、このノイズから元のデータを再構築するようにモデルが学習します。デノイジングネットワークは、初期の時間ステップでは粗い形状や構造（低周波成分）の再構築に集中し、時間ステップが進むにつれて詳細（高周波成分）を徐々に洗練していきます。この挙動は、拡散モデルが粗から細へとデータを生成し、周波数成分と隠れた関係を持つことを示しています。

しかし、この挙動と順方向および逆方向プロセスで使用されるノイズとの関係についての研究は限られています。既存の拡散モデルの多くは、ガウスノイズ（非相関ガウスノイズまたはガウス白色ノイズとしても知られる）にのみ依存しています。このノイズの周波数パワースペクトルはすべての周波数にわたります（白色光に類似）。相関ノイズについては、拡散モデルで十分に検討されていませんが、この領域に関連する研究はいくつか存在します。例えば、Rissanen ら [2023] は、熱拡散に着想を得た拡散プロセスを提案し、周波数を明示的に制御しています。同様に、Voleti ら [2022] は、スコアベース拡散モデルにおいて等方的ガウスノイズの代わりに非等方的ノイズを使用することを提案しています。しかし、これらの手法は生成画像の品質に関する制約に直面しており、主流モデルへの採用が限定的である理由を説明しています。

本研究では、時間変化するノイズを使用した拡散プロセスをサポートする新しい拡散モデルを提案します。本モデルの目標は、ブルーノイズ（[Ulichney 1987]）などの相関ノイズを活用し、生成プロセスを強化することです。ブルーノイズは、低周波領域にエネルギーを持たないパワースペクトルを特徴とします。我々はブルーノイズマスク（[Ulichney 1999]）に注目し、ブルーノイズ特性を持つノイズプロファイルを提供します。これらのブルーノイズマスクを使用して、拡散ベースの生成モデルのための時間変化するノイズを設計することを提案します。ただし、拡散のための相関ノイズマスクを生成するプロセスは時間がかかり、数千から数百万のマスクをその場で生成する必要がある場合があります。この問題に対処するため、我々は低次元および高次元画像の両方に対してリアルタイムでガウスブルーノイズマスクを生成する効率的な方法を提案します。

貢献の概要

- 我々のフレームワークに基づき、時間変化ノイズを使用した決定論的拡散プロセスを導入し、モデルの各ステップで導入される相関を制御可能にします。
- 相関ノイズマスクを生成する際の計算上の課題を克服するため、リアルタイムのマスク生成アプローチを導入します。
- 提案する時間変化ノイズモデルを使用してガウスノイズとガウスブルーノイズを補間することで、既存の決定論的モデル（例：IADB [Heitz et al. 2023]、DDIM [Song et al. 2021a]）を様々な画像生成タスクにおいて上回る性能を達成します。

2 関連研究

2.1 ブルーノイズ

ブルーノイズは、高周波成分を特徴とし、低周波成分を含まないノイズの一種です。コンピュータグラフィックスの分野で数多くの応用が見られます。その一例として、Ulichney [1993] によって初めて提案されたブルーノイズマスクを使用した画像ディザリングが挙

げられます。これにより、画像の知覚品質が向上します。また、ブルーノイズマスクは、Georgiev and Fajardo [2016] および Heitz and Belcour [2019] によって示されたように、モンテカルロレンダリングにおいて誤差分布を改善するためにも利用されています。さらに、ブルーノイズとレンダリングにおけるデノイジングとの関係は、Chizhov et al. [2022] および Salaün et al. [2022] によって探求されており、ブルーノイズをローパスフィルタと組み合わせることで知覚誤差を低減できることが明らかになっています。このようなブルーノイズマスクのデノイジング特性を活用するため、我々はこれを加法的ノイズとして使用し、拡散ベースの生成モデリングにおいてデータを破損させる手法を提案します。

2.2 拡散モデル

拡散モデルによる画像生成には、確率的アプローチ [Ho et al. 2020; Song and Ermon 2019; Song et al. 2021b] と決定論的アプローチ [Song et al. 2021a; Heitz et al. 2023] を含む様々な定式化があります。拡散モデルは画像生成だけでなく、ビデオ生成 [Ho et al. 2022] や 3D コンテンツ生成 [Poole et al. 2023] にも拡張されています。より包括的なレビューについては、Cao et al. [2024] および Po et al. [2023] による調査を参照してください。

拡散モデルは、トレーニングおよび生成プロセスが遅いことが知られています。生成プロセスを高速化し、少ないステップで画像を生成する方法は、ますます重要な研究課題となっています [Lu et al. 2022; Liu et al. 2023a,b; Salimans and Ho 2022; Karras et al. 2022, 2023; Song et al. 2023; Luo et al. 2023]。推論ステップの削減とは別に、一部の研究は、様々な種類のノイズ追加 [Jolicoeur-Martineau et al. 2023] や画像破損操作 [Bansal et al. 2024] をサポートするより一般的なフレームワークの開発に焦点を当てています。さらに、一部の研究では、生成プロセスを粗から細へとモデル化するために、画像コンテンツの周波数を明示的に考慮しています [Rissanen et al. 2023; Phung et al. 2023]。

しかし、画像破損に使用されるノイズの周波数が拡散ベースの生成モデリングにおけるデノイジングプロセスにどのように影響を与えるかを研究した事例は限られています。この問題を理解するため、我々は相関ノイズを利用してデノイジングプロセスを改善するフレームワークを提案します。

3 提案手法

拡散ベースの生成モデルは、前進プロセスと後進プロセスという 2 つの主要なプロセスで構成されます。

3.1 前進プロセス

前進プロセスでは、ノイズ ϵ が導入され、離散時間パラメータ t によって決定されるスケーリング係数を用いて初期画像 x_0 を破損させます。ここで、x_0 はトレーニングデータ分布 p_0 からサンプリングされた実画像を表します。時間ステップ t は 0 から $T-1$ までの範囲を取り、T は離散的な時間ステップの総数です。破損された画像と対応する時間ステップ t はニューラルネットワーク $f_\theta(x_t, t)$ をトレーニングする入力として使用されます。

3.2 後進プロセス

後進プロセスでは、トレーニング済みのネットワークを使用して純粋なノイズをデノイズし、新しい画像を生成します。図 2 にこのプロセスを示します。ガウスノイズ（青い分

布）から始まり、画像はネットワークを介して反復的に処理され、最終的に完全にデノイズされた画像（赤い分布）を得ます。このプロセスの中間ステップでは、ノイズと画像が混ざり合った状態が存在します。図には3つの例が示されています。時間ステップが進む（t が0に近づく）につれて、画像の品質が向上し、より多くの詳細が現れます。この例では、中間ノイズが時間変化スケジュールに従い、ガウスノイズからガウスブルーノイズへと移行します（セクション3.2を参照）。

3.3 相関の検討

本節では、2つの異なる軸、すなわちノイズのピクセル間およびミニバッチ内の画像間の相関を探ります。ノイズマスクと画像間の相関の影響を示すため、Heitz et al. [2023] の研究に基づき、時間変化ノイズを用いた決定論的拡散プロセスを構築しました（IADB手法と呼ばれる）。シンプルさと一対一の比較の公平性を保つため、IADBを基盤として本手法を開発し、新しい要素として説明されるもの以外の特性やハイパーパラメータは保持しました。しかし、本手法は他の既存の生成拡散プロセスにも適用可能な一般的な手法です。

3.4 IADBにおける前進・後進プロセスと目的関数

IADB手法における前進および後進プロセス、さらに目的関数は次のように定義されます：

$$x_t = \alpha_t \epsilon + (1 - \alpha_t) x_0 \tag{1}$$

$$x_{t-1} = x_t + (\alpha_t - \alpha_{t-1}) f_\theta(x_t, t) \tag{2}$$

$$L_{\text{IADB}} = \sum_t (f_\theta(x_t, t) - (x_0 - \epsilon))^2 \tag{3}$$

ここで、x_0 は目標画像、$\epsilon \sim \mathcal{N}(0, I)$ はランダムなガウスノイズ、α_t および α_{t-1} は2つのブレンディング係数を表します。ネットワークモデルは f_θ と呼ばれ、破損された画像 x_t と時間ステップ t を入力として受け取ります。

IADBの確率的定式化も存在しますが、本研究ではその安定性のため決定論的変種に焦点を当てます。

3.5 3.1 相関ノイズ

決定論的な拡散プロセスでは、ノイズマスクが後進プロセスの初期化として画像生成に使用され、また各トレーニングステップで目標画像を破損させるために用いられます。トレーニング中のマスク生成は重要な要素であり、特定の要件を満たす必要があります。このプロセスは確率的である必要があり、各反復で異なるマスクを生成することで、過学習を防ぎ、生成結果の多様性を高めます。また、マスク生成はトレーニングの各ステップで使用されるため、高速であることが求められます。

ガウスノイズはこれらの要件を自然に満たしますが、すべての相関マスク生成手法が同様ではありません。特に、IADBは平均0および単位共分散行列を持つ多変量ガウス分布から生成されたマスクを使用します。一方、ブルーノイズのような相関ノイズを作成するには、非単位共分散行列が必要です。

ブルーノイズマスクの共分散行列 Σ は、マスクのコレクションから推定できます。我々は、Ulichney [1993] の目的関数を使用したシミュレーテッドアニーリングを採用し、1万個のブルーノイズマスクを生成しました。この方法は高品質なマスクを生成しますが、最適化に多大な時間を要します。その後、例となるマスクのそれぞれの共分散行列を平均化することで、ブルーノイズの相関行列 Σ を計算できます。

指定された共分散行列 Σ を持つノイズマスクを作成するには、Σ に対してコレスキー分解を適用し、下三角行列 L を得ます（$LL^\top = \Sigma$）。最後に、ランダムベクトルに L を乗じることで、効率的に目的のノイズマスクを生成できます：

$$b = L\epsilon \tag{4}$$

ここで、$\epsilon \sim \mathcal{N}(0, I)$ は単位分散のガウス分布を表します。図 3 は、式 (4) を使用して生成されたガウスブルーノイズマスクの一例を示しています。

L の各行または列はノイズマスクのピクセルインデックスを表します。L の各セルは、ノイズマスク内のピクセル間の相関強度を示します。正の値は正の相関を表す明るいセルに対応し、負の値は負の相関を表す暗いセルに対応します。各ピクセルについて、隣接するピクセルのみがゼロから離れた値を持ち、それ以外の非隣接ピクセルはゼロに近い値を持つことが、白と黒の線によって示されています。

なお、本論文では b をガウスブルーノイズと呼びますが、Ahmed et al. [2022] の提案する Gaussian Blue Noise 手法とは異なります。

行列 L が高次元である場合、行列-ベクトル積の計算コストが高くなります。高次元ノイズを生成するために直接行列サイズを増やすと、PyTorch [Paszke et al., 2017] のような最新の機械学習フレームワークを使用して（非相関）ガウスノイズを生成する場合に比べて遅くなります。ノイズ生成は各トレーニングステップで使用されるため、このオーバーヘッドは最小限に抑える必要があります。そのため、高次元ノイズマスクを生成するために、Kollig and Keller [2002] の手法を適応し、低次元マスクのセットをパディングしてブルーノイズマスクを作成しました。

具体的には、高次元のガウスブルーノイズマスクのバッチを生成するために、解像度 64^2 でのより大きなバッチを式 (4) を使用して生成します。この場合、$L \in R^{64^2 \times 64^2}$ となります。その後、これらの 64^2 マスクをパディングしてより大きなタイルにし、高次元のガウスブルーノイズマスクを生成します。この方法により、解像度 128^2 のガウスブルーノイズを生成する計算オーバーヘッドはごくわずか（約 0.0002 秒）です。図 1 は、パディングによって生成された解像度 128^2 のガウスブルーノイズの例を示しています。

高次元マスクにパディングを使用すると、パディングされたタイル間にシーム（継ぎ目）が生じますが、このアーティファクトは実際にはほとんど目立たず、手法の低いオーバーヘッドによって補われます。我々は、異なる解像度でのマスクを提供し、Supplemental ドキュメントのセクション 3 において、それぞれのガウスブルーノイズの周波数パワースペクトルを示しています。この結果から、我々のパディング手法を使用することで、異なる解像度においてもブルーノイズの特性が保持されていることが確認されました。

3.6 時間変化ノイズを伴う拡散モデル (Diffusion Model with Time-varying Noise)

単一の行列 L を使用すると、1 種類の相関のみが生成されます。しかし、拡散モデルでは各時間ステップで導入される相関量を制御する必要があります。時間変化する L_t は、2 種類の相関を符号化した 2 つの固定行列を用いて次のように計算されます：

$$L_t = \gamma_t L_w + (1 - \gamma_t) L_b, \tag{5}$$

ここで、L_w と L_b は異なる 2 種類の行列を表し、γ_t はブレンド係数です。これに基づき、フォワードプロセスは次のように定義されます：

$$x_t = \alpha_t (L_t \epsilon) + (1 - \alpha_t) x_0, \tag{6}$$

ここで、このフォワードプロセスにより、時間ステップ t に基づいてガウスノイズとガウスブルーノイズが補間されます。より一般的には、このモデルは任意の 2 種類のノイズを γ_t に基づいて滑らかに補間することをサポートします。

図 4 は、ガウスノイズからガウスブルーノイズへの線形補間の例を示しています。対応する周波数パワースペクトルは離散フーリエ変換によって計算され、低周波領域のエネルギーが左から右へと減少していることが示されています。

次に、フォワードプロセスを逆転させてバックワードプロセスを定義する必要があります。L の定義とフォワードプロセスに基づき、バックワードステップは次のように導出されます：

$$x_{t-1} = x_t + (\alpha_t - \alpha_{t-1})(x_0 - L_t\epsilon) + (\gamma_t - \gamma_{t-1})\alpha_{t-1}(L_b\epsilon - L_w\epsilon). \tag{7}$$

詳細な導出は補足資料のセクション 1 に記載しています。ここで、L_w はガウス（白色）ノイズを表す単位行列であり、L_b は式 (4) で定義された行列です。$L_b = L_w$ の場合、我々のモデルは IADB に戻ります。一方、$L_b \neq L_w$ の場合、時間変化するノイズを伴うより一般的なモデルが得られます。

IADB では、ネットワークは $x_0 - L_t\epsilon$ の項のみを学習するよう設計されていますが、L_t は単に単位行列として扱われています。ここでは、式 (7) の両方の項を学習するようネットワークを訓練します。単純な方法としては、2 つのニューラルネットワークを使用することが考えられますが、これは IADB に比べて計算量が大幅に増加するため現実的ではありません。その代わりに、2 つの項をそれぞれ 3 チャンネルの画像として表現し、合計 6 チャンネルの出力を持つネットワークを設計しました。これらの出力をそれぞれ $f'_\theta(x_t, t)$ と $f''_\theta(x_t, t)$ とし、ネットワークの出力は次のように求められます：

$$f'_\theta(x_t, t) = x_0 - L_t\epsilon, \quad f''_\theta(x_t, t) = \alpha_{t-1}(L_b\epsilon - L_w\epsilon).$$

したがって、損失関数は次のようになります：

$$L_{\text{Ours}} = \sum_t \left((f'_\theta(x_t, t) - (x_0 - L_t\epsilon))^2 + \frac{\gamma_t - \gamma_{t-1}}{\alpha_t - \alpha_{t-1}} (f''_\theta(x_t, t) - \alpha_{t-1}(L_b\epsilon - L_w\epsilon))^2 \right). \tag{8}$$

本モデルは時間変化するノイズで訓練されますが、バックワードプロセス中は依然として決定論的です。バックワードプロセスは初期のガウスノイズから開始し、中間の時間ステップで追加のノイズは必要ありません。代わりに、ネットワークは時間変化するデノイジング手法でバックワードプロセスを誘導するように学習します。

3.7 アルゴリズムおよびノイズスケジューラ (Algorithms and Noise Scheduler)

アルゴリズムの概要　フォワードプロセス、バックワードプロセス、およびノイズ生成の手順をアルゴリズム 1 から 3 にまとめます。アルゴリズム 2 では、`get_alpha`（α-スケジューラ）を Heitz ら [2023] に従い線形関数（$\alpha_t = t/T$）として考えますが、非線形関数にも拡張可能です。

次に、`get_gamma` を式 (9) に基づく一般的なシグモイド関数として定義します。式 (8) における加重項 $(\gamma_t - \gamma_{t-1})/(\alpha_t - \alpha_{t-1})$ は、α-スケジューラと γ-スケジューラの違いを自動的に補正します。$\gamma_t - \gamma_{t-1}$ が小さい場合、$f''_\theta(x_t, t)$ の寄与が減少します。この挙動は、式 (7) に記載されたバックワードプロセスと一致しており、$\gamma_t - \gamma_{t-1}$ が小さい場合、$f''_\theta(x_t, t)$ の重要性が低下します。

ノイズスケジューラ　Chen [2023] の研究に触発され、スケジューラの選択は特に画像解像度が高い場合に重要な影響を及ぼすことが示されています。`get_gamma`（γ-スケジューラ）は、2つのノイズの補間を制御するためにシグモイドベースの関数としてパラメータ化されます。具体的には、γ-スケジューラは Chen [2023] に従い、以下の3つのパラメータで定義されます：start、end、τ。

$$\text{sigmoid}\left(\frac{\text{start} + (\text{end} - \text{start}) \cdot t/T}{\tau}\right), \tag{9}$$

ここで、$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$、$t$ は時間ステップを表します。

パラメータの最適化　start、end、および τ の値を事前に設定する方法が不明であるため、ネットワークパラメータに加えてこれらのパラメータを最適化対象とします。最適化範囲は以下の通りです：$\text{start} \in [-3, 0)$、$\text{end} \in (0, 3]$、$\tau \in [0.01, 1000.0]$。初期実験では、start と end はそれぞれ約 0 と 3 に安定して収束し、τ は画像解像度に応じて約 0.2 に収束するか増加を続ける傾向が確認されました。

しかしながら、これら3つのパラメータを最適化すると、追加のエポックが必要となり、ネットワークの学習が困難になる場合がありました。これは、エポックを通じてパラメータが変化するためです。

パラメータの固定化　これら3つのパラメータの選択をより実用的にするために、$\text{start} = 0$、$\text{end} = 3$ と固定し、τ は画像解像度に基づいて設定しました。具体的には、128^2 解像度では $\tau = 0.2$、64^2 解像度では $\tau = 1000$ と設定しました。

異なる τ 値に対応する γ-スケジューラの曲線を図で示しています。すべての実験で使用した3つのパラメータの値を補足資料セクション2にまとめています。

3.8 考察 (Discussion)

提案した時間変化ノイズモデルは、データに依存した γ_t スケジューラを選択するための柔軟性を提供し、デノイジングプロセスの改善を可能にします。一方で、γ-スケジューラの最適なパラメータを探索するために追加のエポックが必要になる可能性があります。この問題を軽減するため、初期最適化結果に基づいて τ を固定するという実用的な解決策を提案しました。しかし、γ-スケジューラをより効率的に選択する方法については、今後さらなる研究が必要です。

3.9 3.3 整流マッピングを用いたデータサンプルの相関 (Data Sample Correlation Using Rectified Mapping)

これまでの記述では、画素間の相関を利用して拡散プロセスを強化する方法を説明しました。相関は、単一のミニバッチ内でも活用可能であり、ノイズとターゲット画像間のマッピングを改善することができます。

Rectified Flow [Liu et al., 2023a] および Instaflow [Liu et al., 2023b] に触発され、相関を利用してノイズと画像のペアを整流することができます。図5は、トレーニングイテレーション中の単一ミニバッチにおけるデータサンプル x_0（赤い分布）とノイズ b（青い分布）のペアリング、および提案する整流マッピングを視覚化しています。

従来の手法（図5(a)）では、x_0 と b の間にランダムなマッピングを適用していました。本研究では、フォワードプロセスに入力する前に、文脈に応じた層別化（in-context stratification）を適用することで、ノイズデータのマッピングを改善しました（図5(b)）。

この整流マッピングにより、各ノイズとそのターゲット画像間の距離が短縮され、より直接的な軌跡が得られます。

マッピングを見つけるためには、ノイズと画像間の個々の画素レベルでの距離を L2 ノルムを用いて計算します。そして、各 b に対して、これまでに使用されていない x_0 の中から最短距離を持つものを選択します。この改良されたマッピングにより、特定の画像がトレーニングプロセス中に一貫して同じタイプのノイズと関連付けられるようになり、時間ステップを通じて滑らかな勾配フローが得られます。

4 実験 (Experiments)

4.1 4.1 実装の詳細 (Implementation Details)

本研究では、CelebA [Lee et al., 2020]、AFHQ-Cat [Choi et al., 2020]、LSUN-Bedroom [Yu et al., 2015] のデータセットを使用し、さまざまな解像度で無条件/条件付きの画像生成を行いました。実験設定の詳細は付録 Sec. 2 に記載されています。本フレームワークは PyTorch [Paszke et al., 2017] を使用して実装されており、Song et al. [2021a] および Heitz et al. [2023] に基づく公式実装を利用しています。ネットワークには、diffusers ライブラリ [von Platen et al., 2022] に実装された 2D U-Net [Ronneberger et al., 2015] を使用しました。ネットワークアーキテクチャやトレーニングの詳細、Eq. (9) における τ の値などは、付録 Sec. 2 に記載されています。

拡散モデルのハイパーパラメータとして、トレーニングには $T = 1000$、テストには $T = 250$ を使用しました。ネットワークパラメータの最適化には AdamW オプティマイザ [Loshchilov and Hutter, 2017] を学習率 0.0001 で使用しました。すべてのデータセットでのトレーニングとテストには、4 枚の NVIDIA Quadro RTX 8000 (48GB) GPU を用いました。

評価には、FID [Heusel et al., 2017]、Precision および Recall [Kynkäänniemi et al., 2019] を用い、すべてのモデルの生成品質を測定しました。これらのメトリクスは [Stein et al., 2024] の実装を使用し、Inception-v3 ネットワーク [Szegedy et al., 2016] をバックボーンとして用いて計算しました。すべてのデータセットに対し、FID、Precision、Recall を計算するために 30,000 枚の画像を生成しました。

4.2 4.2 画像生成 (Image Generation)

提案手法を、既存の決定論的拡散モデルである DDPM [Ho et al., 2020]、DDIM [Song et al., 2021a]、IADB [Heitz et al., 2023] と比較し、無条件画像生成の性能を評価しました。公平な比較を確保するため、すべての手法において初期ガウスノイズを統一して生成プロセスを進めました。DDIM のトレーニングには diffusers ライブラリを使用し、IADB や本研究と同様の設定でトレーニングを実施しました。また、DDPM や IHDM [Rissanen et al., 2023] などの確率論的拡散モデルとも比較しました。

AFHQ-Cat (64×64)、LSUN-Church (64×64)、CelebA (64×64) のデータセットにおける結果を図 11 に示します。提案手法では、$t = 75$ 頃からブルーノイズ効果が現れ、他の手法とは異なる特徴が視覚的に確認できます。$t = 0$ の時点で生成された画像において、提案手法は建物の柱周辺の歪みが少なく、窓や扉の詳細がより明瞭に描写されていました。

高解像度の結果では、$t = 75$ 頃から生成されたコンテンツに違いが見られます（図 12）。バックワードプロセスの終盤（$t = 25$ 頃）では、ブルーノイズ効果が現れ始めます。この詳細を確認するためには、画像を拡大することを推奨します。また、付録の HTML

ビューアでは、バックワードプロセス中の各時間ステップにおける生成画像を対話的に閲覧できます。

CelebA (128 × 128) データセットに関する定量的評価結果を表 1 に示します。この結果では、DDIM が IADB や提案手法よりも優れた性能を示しました。これは、DDIM が Eq. (1) の α_t に異なる式を採用していることが原因と考えられます。データセットによっては、DDIM が IADB よりも優れた性能を発揮することがありますが、これは IADB や提案手法の限界ではありません。

さらに、トレーニング中に整流マッピングを適用した場合の影響も評価しました。AFHQ-Cat (64 × 64) における拡散ステップ数に対する FID スコアを表 2 に示します。ミニバッチ内でのデータ相関を考慮すると、ステップ数が少ない場合に FID が低下しましたが、ステップ数が多い場合には若干高くなりました。

付録 Sec. 3 では、ガウシアンブルーノイズ生成およびバックワードプロセスの詳細なタイミング、トレーニングデータへの過学習を確認するための最近傍検証を含む追加結果を提供しています。

4.3 他の拡散モデルへの拡張 (Extension to Other Diffusion Models)

本手法は、DDIM への拡張が可能であり、導出および予備的な結果が付録 Sec. 1, 3 に記載されています。また、高解像度画像生成のために LDM [Rombach et al., 2022] に組み込むことも可能です。図 9 に示すように、AFHQ-Cat (512 × 512) において IADB と比較してより現実的な目を生成し、FID (11.45 ¡ 12.19) も改善しています。ただし、他のモデルに基づいて新しいフレームワークを開発するには追加の努力が必要であり、これは将来の課題とします。

4.4 4.3 条件付き画像生成 (Conditional Image Generation)

ノイズからの無条件生成に加えて、本モデルは条件付き画像生成（例: 画像超解像）にも対応しています。具体的には、条件付きの低解像度画像をノイズ画像と連結して入力とするだけで動作します。

図 6 では、LSUN-Church データセットにおける画像超解像（解像度 32 × 32 から 128 × 128 への変換）について、IADB と提案手法の比較を示しています。提案手法は、SSIM [Wang et al., 2004]、PSNR、および平均二乗誤差（MSE）の定量評価で IADB を上回っています。特に、提案手法は MSE が低いことから、参照画像との忠実度において IADB より優れていることが確認されます。

視覚的には、IADB は特に最初の画像の下部で過剰な詳細を導入する傾向があります。一方、提案手法は画像全体で直線を効果的に保持しています。すべての画像超解像実験の定量的結果は、付録 Sec. 3 に記載されており、提案手法が一貫して IADB を上回っていることを示しています。

4.5 4.4 アブレーション研究と分析 (Ablation Study and Analysis)

ノイズの組み合わせ (Combinations of Noises): ガウシアンブルーノイズが高周波特性によって効果的であることを確認するため、ガウシアンブルーノイズを低周波ノイズであるガウシアンレッドノイズに置き換えました。レッドノイズは、同じ手法 [Ulichney, 1993] を用いて、目的関数を最小化する代わりに最大化することで生成されます。その後、対応する共分散行列および下三角行列を計算し、本フレームワークで使用可能なガウシアンレッドノイズを生成しました。

図 7 に示すように、本フレームワークにおいてガウシアンレッドノイズを使用すると、低周波特性のために細部の復元が失敗しました。表 3 では、ガウシアンブルーノイズをガウシアンレッドノイズに置き換えると、Precision が大幅に低下する一方で、Recall はほぼ同等であることが示されています。これは、Precision が主に生成画像のリアリズムを測定する指標であるため、図 7 の視覚的観察と一致しています。

4.6 4.4 アブレーション研究と分析 (続き)

ガウシアンブルーノイズのみの使用: もう一つの選択肢として、ガウシアンブルーノイズのみを使用する方法があります。図 1（第 2 行）に示されている通り、最終生成画像は IADB や提案手法と比較してリアリズムが劣ります。この視覚的な品質は表 3 の定量的な評価結果とも一致しています。

しかし、ガウシアンブルーノイズのみを使用した場合、他の選択肢と比較して、初期の時間ステップで画像の内容がより速く、よりクリーンに現れることが観察されました（図 1 参照）。これを検証するため、早期停止実験を行いました。これは、初期の時間ステップで生成を停止した場合、ガウシアンブルーノイズのみを使用する方がガウシアンノイズのみを使用するよりも良好な結果を得られることを示すものです。

付録 Sec. 3 に記載された図 3 では、ガウシアンブルーノイズのみを使用し、$t = 200$ で停止した結果（第 2 行）が、ガウシアンノイズのみを使用した場合（第 1 行）よりもシャープな詳細を持つことが確認されます。しかし、ブルーノイズには低周波成分が存在しないため、拡散プロセスが限られた方向に制約されます。そのため、中間生成内容を後半の時間ステップで洗練するのが困難となり、最終的に表 3 に示されるような品質の低下を招きます。

提案手法では、中間以降の時間ステップでブルーノイズを取り入れることで、低周波成分がすでに視覚化されている段階で、高周波の詳細を洗練することに焦点を当てています。

異なるノイズの大きさによる拡散: 全ての時間ステップでガウシアンブルーノイズを使用すると品質が劣化するため、後半の時間ステップにおける拡散を分析しました。具体的には、初期の時間ステップを明示的に無視し、特定のノイズの大きさ（例: 30

図 8 に基づくと、ガウシアンブルーノイズを使用することで、ガウシアンノイズを使用した場合と比較して、より多くの詳細と内容が保持された画像を生成できることが示されています。これは、低周波成分が視覚化され始めた段階で、ガウシアンブルーノイズがデノイズに適していることを示しています。この結果は、中間または後半の時間ステップでガウシアンブルーノイズを使用するという本手法の考え方と一致しています。

さらなるアブレーション研究: 我々は、異なる γ 値や γ スケジューラーとしてのコサインベーススケジューラー [Nichol and Dhariwal, 2021] との比較も行いました。また、パディング/タイルで使用される異なるガウシアンブルーノイズマスクサイズについても比較しました。詳細は付録 Sec. 3 をご覧ください。

5 5 結論 (Conclusion)

本研究では、相関ノイズを決定論的生成拡散モデルに組み込む新しい手法を提案しました。本手法は、行列ベースの方法を用いて生成される非相関および相関ノイズマスクの組み合わせを活用しています。

異なるノイズの相関を調査することで、ノイズの特性と生成画像の品質との複雑な関係を明らかにしました。我々の発見によれば、高周波ノイズは詳細を保持するのに効果的ですが、低周波成分の生成には苦労し、一方で低周波ノイズは複雑な詳細の生成を妨げます。

最適な画像品質を達成するために、本研究では時間依存的に異なる種類のノイズを選択的に使用し、それぞれのノイズ成分の強みを活用する方法を提案しました。この手法の有効性を検証するために、よく知られた IADB 手法 [Heitz et al., 2023] と組み合わせて広範な実験を行いました。

トレーニングデータと最適化のハイパーパラメータを一貫して維持することで、さまざまなデータセットで画像品質の大幅な改善を一貫して観察しました。これらの結果は、決定論的拡散モデルの画像生成能力を向上させる上で、本手法の優位性を示しています。

今後の展望 (Future Work)

提案モデルは、生成拡散モデルの効率を向上させるためのノイズパターン設計における新たな研究方向を刺激すると考えています。興味深い今後の課題として、提案モデルを拡張し、2 種類以上のノイズを補間することで、低域通過ノイズや帯域通過ノイズなど、より多様な種類のノイズを取り入れることが挙げられます。これにより、拡散モデルのトレーニングやサンプリング効率をさらに向上させる自由度が増す可能性があります。

さらに、トレーニング中のデータサンプル間の相関を構築するための、より高度な手法を設計することも考えられます。このアプローチは、相関ノイズを使用する場合とは独立して行える方法です。また、本フレームワーク（例: 時間変化ノイズモデル）を、確率的モデル [Ho et al., 2020; Song et al., 2021b] や少数ステップモデル [Karras et al., 2022; Song et al., 2023; Luo et al., 2023] に拡張することも興味深い研究方向です。これにより、提案フレームワークを最新のデノイジング拡散モデルへと一般化することが可能になります。

応用面では、本モデルを 2 次元の無条件および条件付き画像生成でテストしました。今後の課題としては、提案モデルを動画や 3D メッシュなど、他のデータ表現を生成する方向に一般化することが挙げられます。

謝辞 (Acknowledgments)

匿名のレビューアーの詳細かつ建設的なコメントに感謝します。Xingchang Huang は、MPI と Google の共同戦略的パートナーシップである、Saarbrücken Research Center for Visual Computing, Interaction and Artificial Intelligence の支援を受けています。本論文の図 1、8、11、12 で使用した CelebA データセット [Lee et al., 2020] のオリジナル画像は、協定に基づき表示しておりません。これらの図に示されている画像はすべて本手法で生成したものです。

References

[1] Abdalla GM Ahmed, Jing Ren, and Peter Wonka. 2022. Gaussian blue noise. ACM Transactions on Graphics (TOG) 41, 6 (2022), 1–15.

[2] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Cold diffusion: Inverting

arbitrary image transforms without noise. Advances in Neural Information Processing Systems 36 (2024).

[3] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. 2024. A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering (2024).

[4] Ting Chen. 2023. On the importance of noise scheduling for diffusion models. arXiv preprint arXiv:2301.10972 (2023).

[5] Vassillen Chizhov, Iliyan Georgiev, Karol Myszkowski, and Gurprit Singh. 2022. Perceptual Error Optimization for Monte Carlo Rendering. ACM Transactions on Graphics 41, 3, Article 26 (Mar 2022), 17 pages. https://doi.org/10.1145/3504002

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. Stargan v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8188–8197.

[7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems 34 (2021), 8780–8794.

[8] Iliyan Georgiev and Marcos Fajardo. 2016. Blue-Noise Dithered Sampling. In ACM SIGGRAPH 2016 Talks (Anaheim, California) (SIGGRAPH '16). Association for Computing Machinery, New York, NY, USA, Article 35, 1 page. https://doi.org/10.1145/2897839.2927430

[9] Eric Heitz and Laurent Belcour. 2019. Distributing Monte Carlo Errors as a Blue Noise in Screen Space by Permuting Pixel Seeds Between Frames. Computer Graphics Forum (2019). https://doi.org/10.1111/cgf.13778

[10] Eric Heitz, Laurent Belcour, and Thomas Chambon. 2023. Iterative α-(de)blending: A minimalist deterministic diffusion model. In ACM SIGGRAPH 2023 Conference Proceedings. 1–8.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems 30 (2017).

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems 33 (2020), 6840–6851.

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022. Video diffusion models. Advances in Neural Information Processing Systems 35 (2022), 8633–8646.

[14] Alexia Jolicoeur-Martineau, Kilian Fatras, Ke Li, and Tal Kachman. 2023. Diffusion models with location-scale noise. arXiv preprint arXiv:2304.05907 (2023).

[15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems 35 (2022), 26565–26577.

[16] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. 2023. Analyzing and improving the training dynamics of diffusion models. arXiv preprint arXiv:2312.02696 (2023).

[17] Thomas Kollig and Alexander Keller. 2002. Efficient multidimensional sampling. In Computer Graphics Forum, Vol. 21. Wiley Online Library, 557–563.

[18] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems 32 (2019).

[19] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards diverse and interactive facial image manipulation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[20] Xingchao Liu, Chengyue Gong, and Qiang Liu. 2023a. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net. https://openreview.net/pdf?id=XVjTT1nw5z

[21] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. 2023b. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In The Twelfth International Conference on Learning Representations.

[22] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017).

[23] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems 35 (2022), 5775–5787.

[24] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023).

[25] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In International Conference on Machine Learning. PMLR, 8162–8171.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[27] Hao Phung, Quan Dao, and Anh Tran. 2023. Wavelet diffusion models are fast and scalable image generators. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10199–10208.

[28] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T. Barron, Amit H. Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. 2023. State of the Art on Diffusion Models for Visual Computing. arXiv preprint arXiv:2310.07204 (2023).

[29] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net. https://openreview.net/pdf?id=FjNys5c7VyY

[30] Severi Rissanen, Markus Heinonen, and Arno Solin. 2023. Generative Modelling with Inverse Heat Dissipation. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net. https://openreview.net/pdf?id=4PJUBT9f2Ol

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10684–10695.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 234–241.

[33] Corentin Salaün, Iliyan Georgiev, Hans-Peter Seidel, and Gurprit Singh. 2022. Scalable multi-class sampling via filtered sliced optimal transport. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) 41, 6 (2022). https://doi.org/10.1145/3550454.3555484

[34] Tim Salimans and Jonathan Ho. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net. https://openreview.net/forum?id=TIdIXIpzhoI

[35] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning. PMLR, 2256–2265.

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021a. Denoising Diffusion Implicit Models. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net. https://openreview.net/forum?id=St1giarCHLP

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021b. Score-Based Generative Modeling through Stochastic Differential Equations. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net. https://openreview.net/forum?id=PxTIG12RRHS

[38] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency Models. In International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202). PMLR, 32211–32252. https://proceedings.mlr.press/v202/song23a.html

[39] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. 2024. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. Advances in Neural Information Processing Systems 36 (2024).

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2818–2826.

[41] Robert Ulichney. 1987. Digital Halftoning. MIT Press.

[42] Robert Ulichney. 1993. Void-and-cluster method for dither array generation. In Electronic Imaging. https://api.semanticscholar.org/CorpusID:120266955

[43] Robert Ulichney. 1999. The void-and-cluster method for dither array generation. SPIE Milestone Series MS 154 (1999), 183–194.

[44] Vikram Voleti, Christopher Pal, and Adam Oberman. 2022. Score-based denoising diffusion with non-isotropic Gaussian noise models. arXiv preprint arXiv:2210.12254 (2022).

[45] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers

[46] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13, 4 (2004), 600–612.

[47] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. arXiv preprint arXiv:1506.03365 (2015).