2P-08

複数解像度で画像を生成可能な拡散確率モデル

荒川 深映 † 綱島 秀樹 †* 堀田 大地 ††* 森島 繁生 †††

† 早稲田大学 †† 東京大学 ††† 早稲田大学理工学術院総合研究所 (*同一貢献)

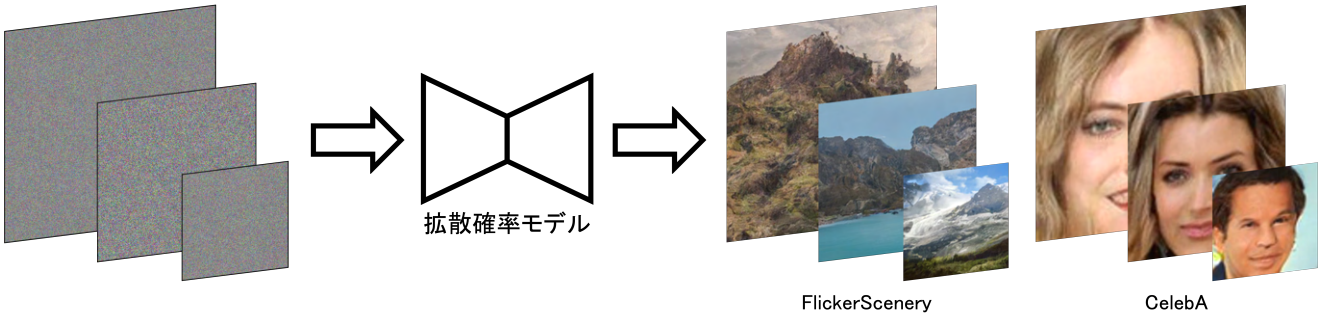

図 1: 提案手法による複数解像度での生成結果．本手法は複数の解像度のノイズから画像を生成できる．

1. はじめに

Stable diffusion (SD) [1] や CLIP [2] のような大規模事前学習モデルは，画像とテキスト間の関係性の学習で著しい成功を示している．特に，SD は拡散確率モデル [3] を利用し，テキスト入力から多様かつ高品質で入力条件に忠実な画像を出力することに成功した．

拡散確率モデルは Generative Adversarial Networks (GANs) を凌駕する程の生成品質や収束性を持つ．しかしながら，学習時に用いた解像度と異なる解像度での推論を行うと，構造が崩壊した画像を出力する問題を持つ．

本研究では，拡散確率モデルが持つ本質的な問題に対処するために，推論時にあらゆる解像度において自然な画像を生成可能な拡散確率モデルの学習方法を提案する．具体的に，我々は 2 つの技術を拡散確率モデルの学習に適用する – (1) ランダムに解像度を変更した画像からパッチを切り取り学習を行うマルチ解像度学習 (2) 解像度の違いをネットワークに伝え条件付けを行う Scale-Aware Feature Adaption (SAFA). 適用するに当たり，特有の条件付けの手法と SAFA の適用法を提案した．

実験では FlickerScenery [4] と CelebA [5] データセットを用いて，既存の拡散確率モデルとの定量的や定性的な評価を通じ，図 1 の通り構造が崩壊することなく複数の解像度での推論を可能にすることを明らかにした．

2. 関連研究

2.1 拡散確率モデル

拡散確率モデル [3] は既知分布からデータ分布への遷移過程を学習することで，画像を生成する手法である．データのスコア関数を U-Net などのネットワークを用いて推定し，デノイジングスコアマッチングによって生成を行う．拡散確率モデルは，多様なデータ分布を捉えられることに加えて，GANs を凌駕する生成品質を持つことから，近年注目を浴びている． しかしながら，学習時に用いた解像度と異なる解像度での推論を行うと，図 2 に示す通り，構造が崩壊した画像を出力する問題を持つ．本研究ではこの問題の解決に取り組む．

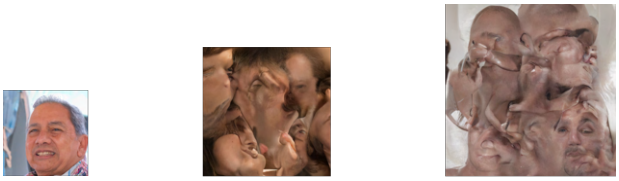

図 2: 256^2 で学習を行った拡散確率モデルでの生成結果

2.2 マルチ解像度学習

マルチ解像度学習 [6] は，GAN で任意解像度の画像生成を行うための手法として提案された．複数解像度の画像を含むデータセット内の画像から，固定解像度のパッチをランダムな位置で切り抜く．切り抜いたパッチに加えて，切り抜いた位置の情報と元画像の解像度を条件として GAN の学習を行うことで，スケールの違いを学習させる．本稿では，このマルチ解像度学習を拡散確率モデルに適用することを考え，条件付け手法を提案する．

2.3 Scale-aware Feature Adaption

Scale-Aware Feature Adaption (SAFA) [7] は複数解像度への超解像手法の主要技術として提案された．一般的な畳み込みでは，入力画像の解像度に問わず同一のカーネルによって処理を行うため，入力画像のスケール変化に対応できない．この問題をカーネルのスケール非依存の問題と呼ぶ．SAFA では，解像度を条件として畳み込みカーネル自体を推定し，推定されたカーネルを用いて畳み込みを行うことで，入力画像の解像度に応じた畳み込みを実現する．本稿では，この技術を適用することで，複数解像度での画像生成を可能としている．

3. 提案手法

3.1 マルチ解像度学習の適用

データセット内の画像の大きさをランダムに変更することで，複数の解像度の画像を含むデータセットを作成し，学習に用いる．拡散ステップ $t \in \{0, 1, \ldots, T\}$ における画像 $\mathbf{x}_t \in \mathbb{R}^{3\times H\times W}$ から，パッチ $\mathbf{p}_t \in \mathbb{R}^{3\times h\times w}$ をランダムな位置で切り抜き，その位置に対応するピクセル座標のグリッド $\mathbf{g} \in \mathbb{R}^{2\times h\times w}$ を計算する． この際の画像の切り抜きと，位置グリッドの対応を図 3 に示す．拡散確率モデルのネットワーク $\boldsymbol{\epsilon}_\theta$ の出力 $\hat{\boldsymbol{\epsilon}}_t$ は，

$$\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_\theta(\mathbf{p}_t, \mathbf{g}, t) \quad (1)$$

Any-resolution Image Synthesis by Diffusion Probabilistic Models: *Shin-ei Arakawa†, Hideki Tsunashima†*, Daichi Horita†*, and Shigeo Morishima††† (†Waseda University, ††The University of Tokyo, †††Waseda Research Institute for Science and Engineering, *equal contribution)*

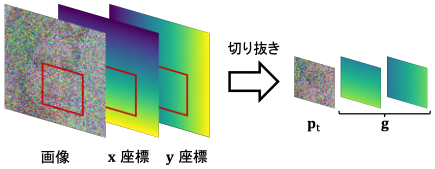

図 3: 画像の切り抜きと位置グリッド $\mathbf{g}$ の対応

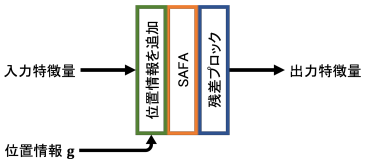

図 4: 提案手法を含めた U-Net の各解像度のブロック

と計算される．位置情報 $\mathbf{g}$ は，フーリエ特徴を用いた変換 [8] を行った後，U-Net を構成する各解像度のブロック内で，特徴量とチャンネル方向に連結することで条件付けを行う．

3.2 Scale-aware Feature Adaption の適用

マルチ解像度学習は画像のスケールの違いをモデルへ学習をさせる目的であったが，2.3 節で述べたスケール非依存の問題は解決されていない．そこで，Scale-Aware Feature Adaption (SAFA) の適用を行う．SAFA は学習可能パラメータによって構成された K 個のカーネル $\{\phi_1, \ldots, \phi_K\} \subset \mathbb{R}^{3\times3}$ をもつ．解像度を入力とするネットワークを用いて，それぞれのカーネルに対する重み $\{w_1, \ldots, w_K\} \subset \mathbb{R}$ を推定する．解像度を考慮した畳み込みカーネル ϕ_{scaled} は，ϕ_k と w_k の線形和 $\phi_{scaled} = \sum_{k=1}^{K} w_k \phi_k$ と定義し，ϕ_{scaled} を用いて畳み込みを行う．

提案手法を含めた U-Net の各解像度のブロックを図 4 に示す．

4. 実験

データセット 実験には FlickerScenery と CelebA データセットを用いた．FlickerScenery は風景画像を集めたデータセットであり，256^2 へ縮小する前処理を行った．CelebA は顔画像を集めたデータセットであり，画像の中心部分で切り抜いて 128^2 の画像を用意した．

学習 マルチ解像度学習では，FlickerScenery と CelebA でそれぞれ最小解像度 s_{min} を 128，64 とし，最大解像度 s_{max} を 256，128 とする．画像は 50% の確率で最小解像度 s_{min} へ縮小し，50% の確率で最小解像度 s_{min} から最大解像度 s_{max} の間の一様分布からサンプルした解像度へ縮小する．切り抜くパッチの大きさは FlickerScenery と CelebA でそれぞれ 128^2，64^2 とする．

解像度の変更 拡散確率モデルの初期ノイズ画像 $\mathbf{x}_T$ の解像度を変更することで，生成画像の解像度を指定する．

4.1 定性評価

FlickerScenery で学習を行ったモデルで，解像度 s_{max} の画像を生成した結果を図 5 に示す．風景の画像には構図としての位置バイアスが含まれていないため，解像度を変更して生成を行ったとしても，不自然に描画されている部分が目立ちにくい．CelebA で学習を行ったモデルでの生成結果を図 6 に示す．高解像度の画像であっても崩れることなく，画像を生成できる一方で，画像の淵付近で特に引き伸ばされた画像となった．

図 5: 最大解像度 s_{max} での生成結果（FlickerScenery）

図 6: 最大解像度 s_{max} での生成結果（CelebA）

表 1: 提案手法での FID(↓) の測定結果

データセット	固定解像度	生成解像度 (提案手法)				
		64^2	96^2	128^2	192^2	256^2
FlickerScenery	19.3	-	-	32.2	64.1	126
CelebA	4.50	11.1	26.8	43.3	-	-

4.2 定量評価

品質と多様性を同時に測定する指標である Fréchet Inception Distance (FID) [9] を用いて，定量評価を行った結果を表 1 に示す．表中の固定解像度とは，それぞれのデータセットで単一の解像度 s_{min} の画像を用いて，既存の拡散確率モデルを学習した結果を指す．固定解像度に比べて提案手法では，解像度 s_{min} をはじめとして，特に高解像度での指標の悪化が顕著であった．本稿では，SAFA をネットワーク内の一部の層にのみ追加を行ったが，依然として残る畳み込み層において，スケール非依存の問題が生じていることが原因だと示唆される．

5. おわりに

本稿では，複数解像度で画像を生成可能な拡散確率モデルを実現するために，マルチ解像度学習と SAFA を拡散確率モデルへ適用するための条件付け手法を提案した．実験から，提案手法によって複数解像度で画像を生成可能であることを示したが，生成品質については改善が必要である．今後は生成品質向上のために，SAFA に代わる複数解像度の畳み込み法について研究を行う予定である．

謝辞

本研究は JSPS 科研費 (19H04137，21H05054) の補助を受けています．

参考文献

[1] R. Rombach *et al.* “High-Resolution Image Synthesis with Latent Diffusion Models”. In *CVPR*, 2022.

[2] R. Alec *et al.* “Learning Transferable Visual Models From Natural Language Supervision”. In *ICML*, 2021.

[3] Ho. Jonathan *et al.* “Denoising Diffusion Probabilistic Models”. In *NeurIPS*, 2020.

[4] C. Yen-Chi *et al.* “In&Out: Diverse Image Outpainting via GAN Inversion”. In *CVPR*, 2022.

[5] L. Ziwei *et al.* “Deep Learning Face Attributes in the Wild”. In *ICCV*, 2015.

[6] C. Lucy *et al.* “Any-resolution training for high-resolution image synthesis”. In *ECCV*, 2022.

[7] L. Wang *et al.* “Learning A Single Network for Scale-Arbitrary Super-Resolution”. In *ICCV*, 2021.

[8] A. Vaswani *et al.* “Attention Is All You Need”. In *NeurIPS*, 2017.

[9] G. Parmar *et al.* “On Aliased Resizing and Surprising Subtleties in GAN Evaluation”. In *CVPR*, 2022.