事前学習済み潜在拡散モデルを用いたゼロショット画像補完
Zero-shot image inpainting using a pre-trained latent diffusion model

柿沼 祐介[†] 宮田 高道[†] 細野 海人[‡] 木下 宏揚[‡]
Yusuke Kakinuma[†] Takamichi Miyata[†] Kaito Hosono[‡] Hirotsugu Kinoshita[‡]
[†] 千葉工業大学 [‡] 神奈川大学
[†]Chiba Institute of Technology [‡]University of Kanagawa

Abstract: ゼロショット画像復元手法である DDNM（Denoising Diffusion Null-space Model）は，事前学習済みの拡散に基づく画像生成モデルを使用しており，タスク固有の学習を行うことなく，さまざまな画像復元タスクに適用することができる．しかし，DDNM は ImageNet で学習された拡散モデルを用いて生成を行なうため，その復元能力は，学習画像のクラス数が限られていることに強く制約される．一方，潜在空間において拡散に基づく生成を行なう latent diffusion models（以下 LDM）は，より大規模なデータセットで学習され，多種多様な画像を生成できることが知られている．LDM を DDNM に適用する際の課題として，DDNM では劣化演算子が線形演算子で表現できることを利用した処理が必要であるのに対し，LDM では前処理として非線形エンコーダを用いていることが挙げられる．本論文では，潜在空間においても原画像の空間的特徴がある程度保持されることに着目し，画像復元タスクを画像補完に絞ることで，より多様な画像の復元を可能とする手法を提案する．実験の結果，提案手法はタスクに特化した学習を行うことなく，多様で高精度な補完が可能であることが示された．

1 はじめに

既存のゼロショット画像復元の一種である Denoising Diffusion Null-space Model (DDNM) [1] は，学習済みのノイズ除去拡散モデルを用いて，画像補完を含む様々な画像復元タスクを解決する．DDNM の課題は，使用する拡散モデルが 1,000 クラスの画像分類データセットである ImageNet で学習されていることである．そのため，DDNM は，ImageNet の 1,000 クラスに含まれない「馬」や「メロン」のような画像を適切に復元できない可能性がある．

これに対し近年，変分オートエンコーダによって得られる低次元の潜在空間における拡散モデルを使用することで，テキストに対応する多様な画像を生成できる Latent Diffusion Models(LDM) [2] ならびにその学習済みモデルの一つである SD（Stable Diffusion） [2] が公開された．以上の背景のもと，本研究では SD と DDNM を適切に組み合わせることにより、高速かつ正確な補完を可能とする画像補完手法を提案する．

2 既存手法

DDNM は，拡散モデルを用いた画像生成技術の一つである DDPM（Denoising Diffusion Probabilistic Models）を利用することで，画像の超解像，白黒画像のカラー化，画像補完，圧縮センシングにおける画像復元，ボケ除去など，(劣化の過程が線形作用素で記述できる) 多様な画像復元を実現する手法である．DDPM は，劣化画像を復元する処理を，現実性の向上および劣化画像との一貫性の維持の 2 つの処理に分割し，それらを交互に適用するアルゴリズムである．なお，拡散モデルの生成過程では，ステップを指す $t \in 0, 1, ..., T$ は T から 0 に向かって進むことに注意されたい．

いま，$\mathbf{x}$ を劣化のない原画像，$\mathbf{A}$ を線形劣化作用素，$\mathbf{y}$ を劣化画像とすると，それらの関係は $\mathbf{y} = \mathbf{A}\mathbf{x}$ と表わせる．DDNM の t ステップ目の処理では，まず現実性を向上させるために，ノイズ除去ネットワークを用いて当該ステップの画像 $\mathbf{x}_t$ から復元画像の推定値である $\mathbf{x}_{0|t}$ を推定する．しかしながら，このステップのみでは得られた $\mathbf{x}_{0|t}$ が一貫性を満たす保証はない．そこで次に，$\mathbf{x}_{0|t}$ を $\mathbf{A}$ の零空間射影した結果から，$\hat{\mathbf{x}}_{0|t} = \mathbf{A}^\dagger\mathbf{A}\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}_{0|t}$ によって一貫性を満たす $\hat{\mathbf{x}}_{0|t}$ を得る．ここで $\mathbf{A}^\dagger$ は $\mathbf{A}$ のムーアペンローズ一般逆行列であり，$\mathbf{I}$ は単位行列である．これにより $\hat{\mathbf{x}}_{0|t}$ は $\mathbf{y} = \mathbf{A}\hat{\mathbf{x}}_{0|t}$ を常に満たす．以上の処理を交互に繰り返すことで，現実性と一貫性の両方を満たす画像を復元できる．

LDM [2] は，学習済みの変分オートエンコーダ (VAE) を用いて画像をより次元の小さい潜在空間上に射影し，その潜在空間上で画像生成処理を行うことで DDPM の学習効率を大きく向上させた画像生成モデルである．LDM の学習済みのモデルの一つである Stable Diffusion（SD）は，50 億枚の画像からなる LAION-5B を学習に使用しており，広い範囲の画像をテキストから生成できる．DDNM の画像生成モデルを学習済み DDPM から SD へと変更することで，復元できる画像のカテゴリを大幅に拡大できることが期待される．

3 提案手法

推定画像の現実性を高めるために，潜在拡散モデルを用いたゼロショット画像補完を提案する．観測モデルは $\mathbf{y} = \mathbf{M}_p \odot D(\mathbf{z})$，であり，ここで D は VAE のデコーダ，$\mathbf{z}$ は原画像と対応する潜在変数であり，$\mathbf{M}_p$ は画素領域の 2 値マスクである．合成演算子 $M_p \odot (D(\cdot))$ の合成は非線形であるため，潜在空間での DDNM の実行は自明ではない．本研究では，潜在空間においても入力画像の空間的特徴がある程度保存されていることに着目し，(画像補完タ

(a) 原画像+マスク　(b) DDNM　(c) 提案手法　(d) 提案手法+BLIP（参考）

図 1: 既存手法と提案手法の定性的な比較.

スクにおける) 潜在空間での零空間射影を次のように書き直すことを提案する.

$$\hat{\mathbf{z}}_{0|t} = \mathbf{M}_l \odot E(\mathbf{y}) + (\mathbf{I} - \mathbf{M}_l) \odot \mathbf{z}_{0|t} \tag{1}$$

ここで E は VAE のエンコーダ，$\mathbf{z}_{0|t}$ は潜在空間拡散過程におけるステップで復元された潜在変数であり，$\odot$ は要素毎の乗算を，$\mathbf{M}_l$ は潜在空間のマスクを表わす．潜在空間のマスク M_l は，M_p に最近傍補間によるダウンサンプリングを適用した後，マスク境界を 2 ピクセル拡張することによって得られる．提案手法では，$\hat{\mathbf{z}}_{0|t}$ から $t-1$ ステップの潜在変数 $\mathbf{z}_{t-1}$ を得る処理は，DDIM と同じものを用いる．以上の処理を $t=0$ になるまで繰り返し，潜在空間変数 $\mathbf{z}_0$ に D を適用することで，補完画像 $\mathbf{x}_0$ が得られる.

4　実験

設定　提案手法の性能を評価するための比較手法として，DDNM [1] を使用した．テスト用のデータセットとしては，BSDS500，ImageNet1k，ImageNet-O，のそれぞれから 200，1000，1000 枚の画像を選んだものを用いた．すべての画像は短辺が 512 ピクセルとなるようにリサイズされ，その中央に 168×168 ピクセルおよび 136×136 ピクセルの欠損領域を設定した．評価指標をしては Fréchet inception distance (FID) を用いた.

結果　表 1 は，既存手法と提案法による原画像と補完画像の間の FID である．画素空間マスク $\mathbf{M}_p$ として，168×168 画素と 136×136 画素の 2 つのセンターマスクを用いる．この表から，提案手法は全てのデータセットとマスクサイズにおいて，最先端のゼロショット画像補完手法の一つである DDNM を凌駕していることがわかる.

提案手法の画像 1 枚あたりの推論時間は，NVIDIA RTX 4080 GPU1 台で約 3 秒であるのに対し，DDNM は約 1 分 40 秒を要する．このことから，提案手法は DDNM に比べて約 33 倍の大幅な高速化を実現している.

表 1: 提案手法と既存のゼロショット画像補完アルゴリズムの FID．値が低い程性能が優れていることを示す．最も優れた結果は**太字**で示す.

データセット	BSDS500		ImageNet-1k		ImageNet-O	
マスクサイズ	168	136	168	136	168	136
DDNM [1]	59.54	43.83	35.63	22.82	36.80	25.25
Ours	**56.57**	**39.90**	**29.43**	**20.49**	**29.05**	**20.40**
Ours+BLIP* (refs. only)	51.05	37.33	27.82	21.18	32.20	23.54

提案手法の特徴の一つである，テキストプロンプトによる補完の精度向上効果を評価するため，入力画像からキャプションを推定する手法である Bootstrapping Language-Image Pre-training (BLIP) [3] を原画像に適用することで得られた理想的なプロンプトを用いた提案手法の FID 値を表 1 に Ours+BLIP*として示す．この結果から，理想的なプロンプトを用いることで補完された画像は，他の手法よりも原画像に近いことがわかる.

ベースライン手法と提案手法による補完画像の視覚的比較を図 1 に示す．これらの画像は，提案手法が現実的な内容で欠損領域を補完できることを示している．この例では，DDNM による補完結果は非現実的であり，欠落領域にアーティファクトが含まれていることがわかる.

5　結論

潜在空間における潜在拡散モデルと零空間射影を組み合わせたゼロショット画像補完手法を提案した．実験結果より，提案手法は定量的・定性的評価において既存手法を上回った．また，潜在空間での処理により，従来の手法と比較して約 33 倍の高速化を実現した．本手法の性能は，原画像から得られた理想的なプロンプトを用いるとさらに向上する．今後の課題として，部分的にマスクされた画像から適切なプロンプトを取得する方法を検討する.

参考文献

[1] Y. Wang *et al.*, "Zero-shot image restoration using denoising diffusion null-space model," in *Proc. of ICLR*, 2023.

[2] R. Rombach *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proc. of CVPR*, 2022.

[3] J. Li *et al.*, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. of ICML*, 2022.

千葉工業大学 先進工学部 知能メディア工学科
〒*275–0016* 千葉県習志野市津田沼 *2-17-1*
Phone: 047-478-0136
E-mail: takamichi.miyata@it-chiba.ac.jp