

# 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<sup>†</sup> 宮田 高道<sup>†</sup> 細野 海人<sup>‡</sup> 木下 宏揚<sup>‡</sup>  
Yusuke Kakinuma<sup>†</sup> Takamichi Miyata<sup>†</sup> Kaito Hosono<sup>‡</sup> Hirotugu Kinoshita<sup>‡</sup>  
<sup>†</sup> 千葉工業大学 <sup>‡</sup> 神奈川大学

<sup>†</sup>Chiba Institute of Technology <sup>‡</sup>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ます。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 latent diffusion models (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ーを用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保持されることに着目し、画像復元タスクとしてに着目することで、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

画像の補完は、入力画像の欠落した領域や特定の領域を視覚的に首尾一貫した内容で埋めることを目的としており、美術品のデジタル修復や写真からの不要なオブジェクトの除去など、幅広い用途がある。

既存の画像補完手法 [1, 2, 3, 4] のほとんどは、現実的でシームレスな補完画像を生成するのに有効であることが証明されている generative adversarial networks (GAN) [5] に依存している。

GAN に代わる有望な選択肢として、ノイズ除去拡散モデルを画像補完 [6, 7, 8] や一般的な画像復元タスク（画像を含む） [9, 10] に適用することが、GAN と比較してより首尾一貫した自然な画像を生成する能力により、大きな人気を集めている。

Wang らは、Denoising Diffusion Null-space Model (DDNM) を提案し、Denoising Diffusion Model 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、DDNM で使用されるノイズ除去拡散モデルが、1,000 クラスの画像分類データセットである ImageNet データセットで学習されることである。したがって、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補完できない可能性がある。

近年、Latent Diffusion Models (LDM) [11] が提案され、variational autoencoder (VAE) 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、より多様で現実的な画像を生成することが可能になった。SD (Stable Diffusion) [11] は LDM の公開モデルの一つであり、50 億枚の画像からなる大規模データセットで学習され、多種多様なテキストプロンプトから知覚的に矛盾のない画像を生成できていることが知られている。しかし、SD における VAE のエンコーダーは非線形であるため、SD を DDNM の基幹と

して直接利用することは困難である。

本論文では、DDNM と SD を組み合わせたゼロショット画像補完手法を提案する。我々は、潜在空間においても原画像の空間的特徴が保存されるという事実に着目する。この重要な観察に基づき、逆拡散において、潜在空間変数上の既知の（欠損していない）領域を、与えられた領域の潜在空間表現で上書きすることを提案する。実験結果は、提案手法がタスク固有の訓練なしに、高速かつ正確な補完が可能であることを示している。

## 2 準備

### 2.1 Denoising Diffusion Models for Image Generation

Wang らは、Denoising Diffusion Null-space Model (DDNM) [9] を提案し、Denoising Diffusion Model 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、DDNM で使用されるノイズ除去拡散モデルが、1,000 クラスの画像分類データセットである ImageNet データセットで学習されることである。したがって、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補完できない可能性がある。

DDPM (Denoising Diffusion Probabilistic Models) [12] は、画像生成拡散モデルの代表的な手法である [12, 13, 14]。DDPM の拡散（順方向）過程は、各ステップが徐々に画像にノイズを付加していくマルコフ連鎖として記述でき、以下のように表される。

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim N(0, \mathbf{I}) \quad (1)$$

ここで、 $t$  は全体の  $T$  ステップの中の特定のステップを指す。さらに、ステップ  $t$  におけるノイズレベルはそれぞれ、 $\alpha_t$  と  $\bar{\alpha}_t = \prod_{s=1}^T \alpha_s$  であり、また、 $\epsilon$  はガウシアンノイズである。

拡散モデルは、ステップ  $t$  で原画像に付加されたノイズ  $\epsilon$  を推定する学習済みニューラルネットワーク  $\epsilon_\theta(\mathbf{x}_t, t)$  を用いて、ステップ  $T$  から 0 までの逆拡散（逆方向）過程を反復することで画像生成を実現する。DDPM の拡張版である DDIM (Denoising Diffusion Implicit Models) [13] の逆拡

散過程を以下に示す.

$$\begin{aligned} \mathbf{x}_{t-1} = & \sqrt{\bar{\alpha}_{t-1}} \left( \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) \right) \\ & + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2(\eta)} \epsilon_\theta(\mathbf{x}_t, t) + \sigma_t(\eta) \epsilon \end{aligned} \quad (2)$$

ここで,  $\sigma_t(\eta) = \eta \sqrt{\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}} \sqrt{\frac{1 - \bar{\alpha}_t}{\bar{\alpha}_{t-1}}}$ . ただし,  $\eta = 1$  の場合, 式 (2) は DDPM の逆過程に等しくなり,  $\eta = 0$  の場合は完全に決定論的なプロセスとなる。

## 2.2 Denoising Diffusion Null-space Model (DDNM)

DDNM は拡散モデルに基づく画像復元手法であり, 画像の超解像, 白黒画像のカラー化, 画像補完を実現する. ノイズ無しの場合, DDDNM の観測モデルは  $\mathbf{y} = \mathbf{Ax}^*$  であり,  $\mathbf{x}^*, \mathbf{A}, \mathbf{y}$  はそれぞれ原画像, 線形劣化演算子, 劣化画像である.  $\mathbf{A}$  を変えることで, 様々な画像復元タスクの観測モデルを定式化できる.

DDNM の要点は, 画像復元を, 推定画像の実在性の向上と劣化画像との整合性の維持の 2 つのステップに分け, 両者を交互に行うことである. DDDNM はまず, 次式に示すように, 事前に学習させた拡散モデルを用いて推定画像の実在性を向上させる.

$$\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) \quad (3)$$

ここで,  $\mathbf{x}_t$  と  $\mathbf{x}_{0|t}$  はそれぞれ  $t$  番目のステップにおける推定ノイズ画像と原画像である. この処理により推定画像の実在性は向上するが, 得られた  $\mathbf{x}_t$  は観測画像  $\mathbf{y}$  との整合性が保証されない. 観測画像  $\mathbf{y}$  との整合性を保つために, DDDNM では, 次式に示すように, 零空間射影と呼ばれる処理を適用する.

$$\hat{\mathbf{x}}_{0|t} = \mathbf{A}^\dagger \mathbf{A} \mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}) \mathbf{x}_{0|t} \quad (4)$$

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  $\mathbf{y}$  と一致する, すなわち  $\mathbf{y} = \mathbf{A} \hat{\mathbf{x}}_{0|t}$  が常に満たされる.  $\mathbf{x}_{t-1}$  は式 (2) が得られる. 以上の処理を交互に繰り返すことで, 実在性と一貫性の 2 つの性質を満たす画像を復元できる.

## 2.3 Latent Diffusion Models (LDM)

DDPM は高次元画素空間で直接動作するため, これらのモデルの学習と推論には大量の計算リソースが必要となる. このリソース消費のため, DDPM の一般に利用可能な事前学習済みモデルは, せいぜい 1,000 クラスの画像分類データセットである ImageNet 上での学習によってのみ得られる. ImageNet には「馬」のような非常に一般的なクラ

---

### Algorithm 1 提案手法のサンプリング

---

```

1:  $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2: for  $t = T$  to 1 do
3:    $\mathbf{z}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{z}_t, t))$ 
4:    $\hat{\mathbf{z}}_{0|t} = \mathbf{M}_l \odot E(\mathbf{y}) + (\mathbf{I} - \mathbf{M}_l) \odot \mathbf{z}_{0|t}$ 
5:    $\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{z}}_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2(\eta)} \epsilon_\theta(\mathbf{z}_t, t) + \sigma_t(\eta) \epsilon$ 
6: end for
7:  $\mathbf{x}_0 = D(\mathbf{z}_0)$ 
8: return  $\mathbf{x}_0$ 

```

---

スが含まれていないため, DDPM を画像生成として用いる DDPM や DDDNM の画像生成や画像復元能力が強く制限される.

この問題に対処するために, Rombach らは, 事前に学習された変分オートエンコーダ (VAE) によって得られる低次元の潜在空間において拡散に基づく画像生成する潜在拡散モデル (LDM) [11] を提案した. 潜在空間における拡散のアイデアにより, LDM は比較的小さな計算資源で学習することが可能となった. Stable Diffusion (SD) [11] は, LDM の事前学習済みモデルで, LAION-5B データセット [15] を学習に使用する. LAION-5B は 50 億枚のキャプション付き画像から構成され, 幅広い画像カテゴリをカバーしている. また, SD は, 画像生成の条件付けに CLIP (Contrastive Language-Image Pre-training) [16] を用いている. このように, DDDNM の画像生成を DDPM から SD に変更することで, 復元画像のカテゴリが大幅に拡大することが期待できる.

しかし, SD を DDDNM に適用するのは簡単ではない.  $z^*$  を原画像  $\mathbf{x}^*$  に対応する潜在空間変数とすると, 観測モデルは  $\mathbf{y} = \mathbf{A}(D(z^*))$  と定式化でき, ここで  $D$  は VAE のデコーダである. この場合, 劣化演算子  $\mathbf{A}$  と VAE デコーダ  $D$  の合成はもはや線形演算子ではなくなるため, 潜在空間において式 (4) と同様の零空間射影をすることは困難となる.

## 3 提案手法

推定画像の実在性を高めるために, 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  $\mathbf{y} = \mathbf{M}_p \odot D(\mathbf{z}^*)$ ,  $\mathbf{M}_p$  は画素領域の 2 値マスクである. 上述したように, 合成演算子  $\mathbf{M}_p \odot (D(\cdot))$  の合成は非線形であるため, 潜在空間での DDDNM の実行は些細なことではない. しかし, VAE エンコーダによって画像が潜在空間に射影された後でも, 原画像の空間的特徴の大部分は保存されるので, 潜在空間における画像補完の場合, 零空間射影を次の

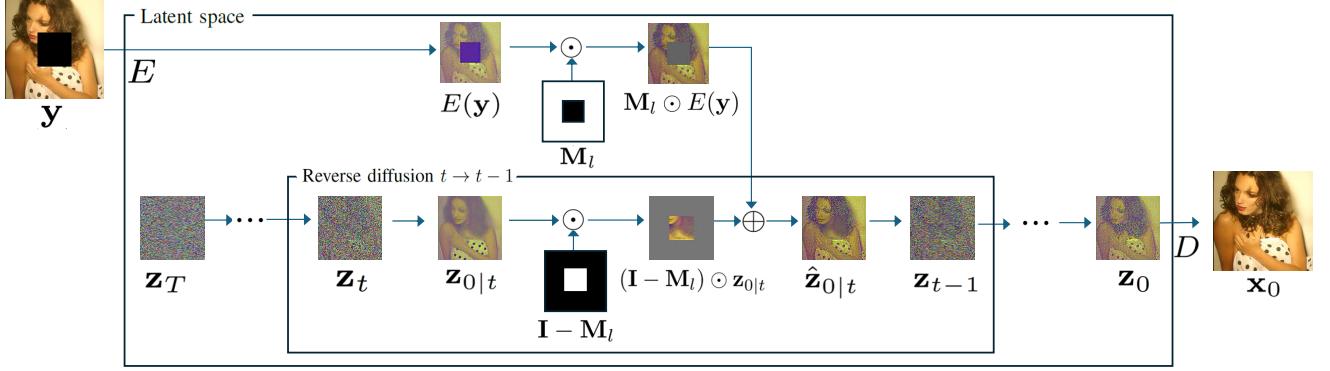


図 1: 提案手法の処理の流れ.

ように書き直せることができた.

$$\hat{\mathbf{z}}_{0|t} = \mathbf{M}_l \odot E(\mathbf{y}) + (\mathbf{I} - \mathbf{M}_l) \odot \mathbf{z}_{0|t} \quad (5)$$

$\mathbf{z}_{0|t}$  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、 $\odot$  は要素毎の乗算を表す。 $\hat{\mathbf{z}}_{0|t}$  は、観測値  $\mathbf{y}$  と一致する調整された潜在変数であり、 $\mathbf{M}_l$  は潜在空間のマスクを表わす。

潜在空間のマスク  $M_l$  を得るために、まず  $M_p$  に最近傍補間によるダウンサンプリングを適用する。次に、マスク境界を 2 ピクセル拡張し、補間領域と非補間領域の境界を平滑化して  $M_l$  を得る。

$\hat{\mathbf{z}}_{0|t}$  から、式 (2) に示す DDIM と同様の処理により、後続の  $t-1$  ステップの潜在変数を以下のように推定できる。

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{z}}_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2(\eta)} \epsilon_\theta(\mathbf{z}_t, t) + \sigma_t(\eta) \epsilon_z, \quad (6)$$

$\epsilon_z$  はランダムなガウシアンノイズで、 $\eta$  は補完アルゴリズムの確率的挙動を制御するパラメータである。 $\eta$  が大きいほど、補完画像はより多様になる。

最後に、潜在空間変数  $z_0$  にデコーダ  $D$  を適用すると、補完画像  $\mathbf{x}_0$  が得られる。提案手法の画像補完アルゴリズム全体をアルゴリズム 1 と画像 1 に示す。

我々の提案手法は、Differential Diffusion (DD) [17] と密接に関連しているが、主な違いは、DD が画像編集タスクに焦点を当てていることである。DD の潜在空間マスクは 2 値ではなく、逆拡散処理中に推定する画素値を徐々に増加させることで、与えられた領域と推定領域の境界を平滑化する。DD を画像補完に適用できるが、既知領域の画素値も変化してしまうため、観察画像と矛盾した補完画像になってしまう。

## 4 実験

提案手法を定量的・定性的に評価するための実験をした。

### 4.1 準備

**比較対象.** 定性的評価では、DeepFill v2 [2]、LaMa [4]、DDNM [9] を比較手法として使用し、定量的評価では、DDNM を比較手法とした。DDNM と我々の提案手法はゼロショットであるが、DeepFill v2 と LaMa はタスク固有のトレーニングが必要であることに注意。

**実装の詳細.** 提案手法の LDM の実装として Stable Diffusion v1.4 [11] を用いた。時間ステップ  $T$  を 30、 $\eta$  を 0.85 に設定し、他のパラメータはデフォルト値に設定した。

DeepFill v2 以外のベースライン手法は、公式実装を使用。DeepFill v2 には非公式の PyTorch 実装 [18] を採用。我々は、DeepFill v2 と LaMa の事前学習済みモデルを使用し、LaMa は Places-2 データセット [19] で学習済みである。オリジナルの DDNM は  $256 \times 256$  ピクセルの画像にしか対応していないため、DDNM の一般化版 [20] に相当し、任意の解像度の画像に適用できる「hq demo」コードを採用した。DDNM のバックボーンとして、ImageNet-1k [21] を用いてクラス無条件設定で学習したガイド拡散モデル [14] を選択した。DDNM のパラメータは、推論を高速化するために時間ステップ  $T$  を 25 に設定した以外は、公式実装のデフォルト値に設定した。

**データセット.** ベンチマークには 3 つのデータセットを用いた。Berkeley Segmentation Data Set (BSDS500) [22]、ImageNet1k O [23]。BSDS500 データセットでは、テストセットの 200 画像すべてを使用し、ImageNet-1k では、検証セットの 1,000 クラスから 1 画像ずつサンプリングした。ImageNet-O では、200 クラスからそれぞれ 5 枚の画像をサンプリングした。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、512 × 512 ピクセルの領域はリサイズされた画像の中心から切り取られた。ImageNet-O は ImageNet-1k データセットに含まれていないクラスの画像から構成されていることに注意が必要である。DDNM のバックボーンである拡散モデルは ImageNet-1k で学習されているため、BSDS500 や ImageNet-O に対する補完性能が低下する可能性がある。

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します。最も優れた結果は**太字**で示されています。

データセット	BSDS500		ImageNet-1k		ImageNet-O	
マスクサイズ	168	136	168	136	168	136
DDNM [9]	59.54	43.83	35.63	22.82	36.80	25.25
<b>Ours</b>	<b>56.57</b>	<b>39.90</b>	<b>29.43</b>	<b>20.49</b>	<b>29.05</b>	<b>20.40</b>
Ours+BLIP* (refs. only)	51.05	37.33	27.82	21.18	32.20	23.54

**評価指標.** Fréchet inception distance (FID) [24] を用いて, 補完画像の品質を評価する.

## 4.2 定量評価

表 1 は, 既存手法と本提案法による, 原画像と補完画像の FID 値である. 画素空間マスク  $M_p$  として,  $168 \times 168$  画素と  $136 \times 136$  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.

提案手法の推論時間は, NVIDIA RTX 4080 GPU1 台で 1 画像あたり約 3 秒であるのに対し, DDNM は同じ GPU で 1 画像あたり約 1 分 40 秒かかる. つまり, 提案手法は DDNM に比べて大幅な高速化 (約 33 倍) を達成している.

提案手法の特徴の一つは, テキストによるプロンプトを利用して, 補完プロセスをガイドできることである. そのため, 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) [25] を原画像に適用し, 「理想的な」テキストプロンプトを得る. この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, 他の手法よりも現画像に近い.

ImageNet-O は, ImageNet1k で学習した画像分類ネットワークが誤認識を引き起こすように意図的に選択された画像で構成されており, BLIP の基幹である CLIP のゼロショット画像認識性能も低いことが知られている [26]. このようなデータセットの特徴が, BLIP のキャプション予測能力, および予測されたキャプションを用いた提案手法のインペイント能力に悪影響を及ぼしていると考えられる.

## 4.3 定性評価

ベースライン手法と我々の提案手法による補完画像の視覚的比較を図 2 に示す. 1 行目と 2 行目の画像は, 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. また, Places-2 データセットで学習した Deepfill v2 と

LaMa も, 1 行目の建物画像の補完に成功している. しかし, 2 行目と 3 行目の画像では, Deepfill v2 と LaMa による塗りつぶし画像は不鮮明な結果となり, 入力画像と一致したコンテンツを生成できない. すべての例において, DDNM による補完結果は非現実的であり, 欠損領域にアーティファクトが含まれている.

## 5 結論

我々は, 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ことさらに向上する. 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.

## Acknowledgment

本研究は, 財団法人電気通信普及財団の研究助成, および日本学術振興会科学研究費補助金 JP23K03871 の一部を受けた.

## 6 参考文献

### 参考文献

- [1] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and locally consistent image completion,” *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 107:1–107:14, 2017.
- [2] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, “Generative image inpainting with contextual attention,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [3] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, “Free-form image inpainting with gated convolution,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4471–4480, 2019.
- [4] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, “Resolution-robust

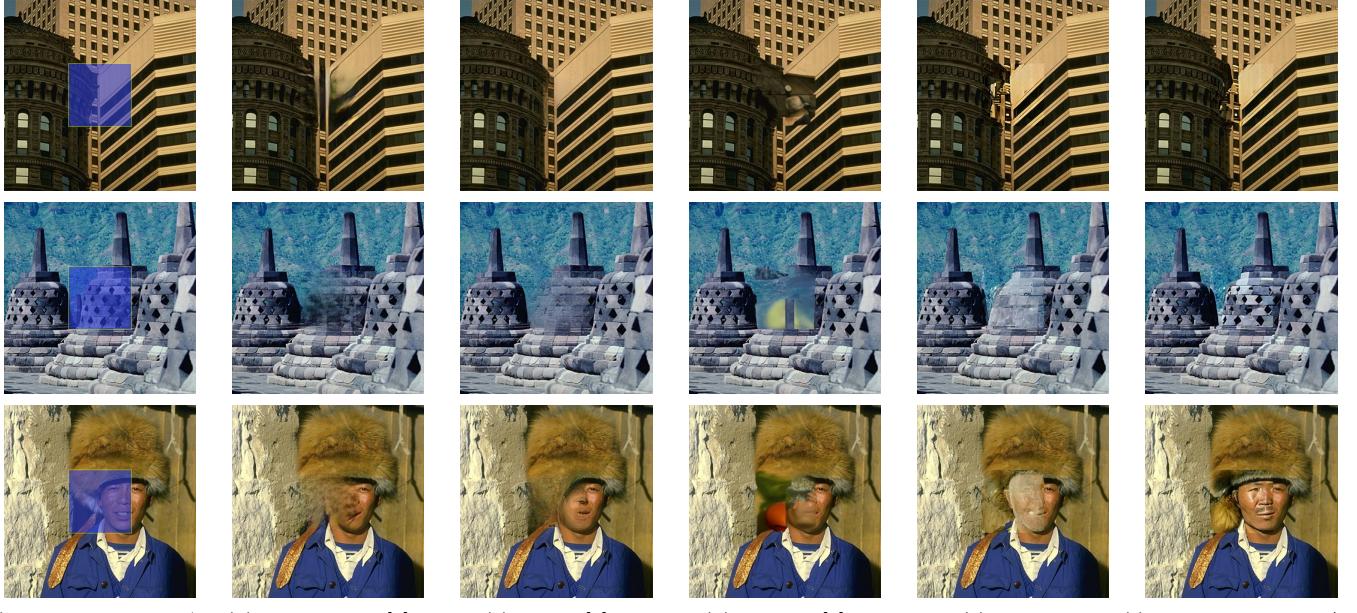


図 2: 既存手法と提案手法の定性的な比較. 入力画像 (a) は, BSDS500 データセットに含まれるオリジナル画像の一部で, 168×168 をマスク済み

large mask inpainting with Fourier convolutions,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2149–2159, 2022.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing System (NeurIPS)*, vol. 27, 2014.

[6] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, “Repaint: Inpainting using denoising diffusion probabilistic models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] O. Avrahami, D. Lischinski, and O. Fried, “Blended diffusion for text-driven editing of natural images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18208–18218, 2022.

[8] S. Xie, Z. Zhang, Z. Lin, T. Hinz, and K. Zhang, “Smartbrush: Text and shape guided object inpainting with diffusion model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[9] Y. Wang, J. Yu, and J. Zhang, “Zero-shot image restoration using denoising diffusion null-space model,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[10] B. Kawar, M. Elad, S. Ermon, and J. Song, “Denoising diffusion restoration models,” in *Proceedings of the ICLR Workshop on Deep Generative Models for Highly Structured Data (ICLRW)*, 2022.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

[12] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[13] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[14] P. Dhariwal and A. Q. Nichol, “Diffusion models beat GANs on image synthesis,” in *Advances in Neural Information Processing Systems* (A. Beygelz-

- imer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021.
- [15] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, “LAION-5B: An open large-scale dataset for training next generation image-text models,” in *Proceedings of the Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [16] 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,” in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 139, pp. 8748–8763, 2021.
- [17] E. Levin and O. Fried, “Differential diffusion: Giving each pixel its strength,” *CoRR*, vol. abs/2306.00950, 2023.
- [18] Y. Zhao, “Deepfillv2.” <https://github.com/zhaoyuzhi/deepfillv2>.
- [19] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.
- [20] Y. Wang, J. Yu, R. Yu, and J. Zhang, “Unlimited-size diffusion restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1160–1167, 2023.
- [21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.
- [22] D. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, vol. 2, pp. 416–423, 2001.
- [23] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, “Natural adversarial examples,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, 2021.
- [24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [25] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 12888–12900, 2022.
- [26] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, “Long-CLIP: Unlocking the long-text capability of CLIP,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.