表現条件付き潜在拡散モデルと表現学習

浮田 嵩祐[1,a)]　YE Xiaolong[1,b)]　大北 剛[1,c)]

概要：本論文では，VAE の潜在空間上で画像表現をエンコードする機構を Transformer ベースの拡散モデル内に組み込み，表現で条件づけられた画像生成モデルを提案する．さらに，クラスラベルで条件づけられた潜在拡散モデルからのゼロショット分類手法を用いて脳画像の血腫分類を行った．これは，強力な対比学習手法である DINOv2 での表現を用いた線形分類の結果を上回る精度が得られた．

Representation Conditional Latent Diffusion Model with Representation Learning

Abstract: In this paper, we propose an image generation model that is conditioned by the representation encoding image representations on the latent space of VAE using a Transformer-based diffusion model. Furthermore, we performed hematoma classification in brain images using a zero-shot classification method based on a class labels conditional latent diffusion model. The accuracy exceeded the results of linear classification using DINOv2 representation.

1. はじめに

画像生成は GAN[1] が提案されて以降, 急速に発展し, 描画する画像の大きさと精細を大幅に上げてきた．しかし, GAN は、モード崩壊やトレーニングの不安定性という根本的な困難さをもつ．このため, 生成方法が異なる別の生成モデルが生み出された．自然言語に対する最も実用的な生成モデルとして開発の進む Transformer から転用されたビジョントランスフォーマーをベースとした画像生成モデル ([2], [3]), 尤度を考慮可能な生成モデルであるフローモデル [4], 段階的なノイズ除去プロセスを介してデータサンプルを再構築する拡散モデル [5] である．VAE による潜在空間上の拡散モデルとして構築された Stable Diffusion[6] は，優れた計算効率と高精細な画像生成を実現して, 大きな印象を与えた．

さらに、ここ数年は, 大規模化に耐えうる手法という観点での研究も急速に発展し, ここでも GAN から拡散モデルへの移行が見られる．1 つ目は, テキスト条件付き (text-to-image) という画像生成法に対する一連の研究である．VQ-VAE[7] は, テキストトークンのシークエンスとそれに続く画像トークンのシークエンスで自己回帰 Transformer をトレーニングする．DALL-E[8] は，CLIP[9] を用いることで，出力画像をランク付けしてフィルタリングし，よりキャプションに沿った画像を生成する．DALL-E 2[10] は，拡散モデルを用いて，テキスト入力を処理する補助テキストエンコーダを使用して拡散モデルをトレーニングする．CLIP によるテキスト埋め込みを拡散モデルで条件付ける．2 つ目は, 画像表現を条件 (image-to-image) とした拡散モデルである．Bordes et al. [11] は対比学習モデルからの画像表現を条件とする拡散モデルを提案した．カーネル密度推定を使用し画像表現をサンプリングすることにより，2 段階の画像生成を行い, 生成された表現を拡散モデルに与えることで，画像をエンドツーエンドで生成可能にした．Jeremias[12] は, 学習済みの画像表現だけでなく, 拡散モデル内でエンコードした表現を与えて訓練する表現条件付き拡散モデルを考案する．Diffusion Transformers(DiTs)[13] は Transformer に基づく拡散モデルに焦点をあてた．

本論文で考えるのは, このような生成モデルで学習された表現を用いて, 画像認識する技術である．画像を, 自己教師あり学習によりラベルなし画像を自己回帰 Transformer で表現を学習して, 後続タスクにより画像認識する形を意識する．この形において, 対比学習の効果が大きいことが

[1] 九州工業大学大学院情報工学研究院知能情報工学研究系大北研究室
[a)] ukita.kosuke299@mail.kyutech.jp
[b)] ye.xiaolong713@mail.kyutech.jp
[c)] tsuyoshi@ai.kyutech.ac.jp

示された．MoCo[14] や SimCLR[15] では, 表現空間内で類似のインスタンスが互いに近くに配置され，異なるインスタンスが遠く離れている表現を学習する．BYOL[16] はこの欠点を改良して, 画像を区別することなく教師なしで特徴を学習する．データ拡張を用いた非対称的な入力と，stop-gradient を用いた非対称的な重み更新による 2 つの非対称性を生み出し，2 つのネットワークの出力の類似度を小さくするように訓練を行う．DINO[3] は入力画像に対して 2 つの異なるランダム変換を施し，生徒と教師のネットワーク両方に渡す．生徒と教師のネットワークの両方ともアーキテクチャは同じだが，パラメータは異なり，教師ネットワークの出力はバッチ全体で計算された平均値を中心に配置される．DINO v2[17] は SwAV[18] を中心として DINO と iBOT[19] を組み合わせる．DINO アルゴリズムに加えて生徒ネットワークへ入力する画像パッチの一部をマスクする点が改良された．

本論文では, 画像表現を条件とする, 表現条件つきの Transformer に基づく潜在拡散モデルを提案する．GAN や自己回帰 Transformer に基づく手法と比較して, 拡散モデルは画像生成において高精細な画像生成を行うために, 表現の精度が高いと考える理由に基づく．

2. 我々の提案

第 2 章では，潜在空間上で動作する Transformer[20] を用いた拡散モデル（DiT[13]）をバックボーンアーキテクチャとして，DiT の詳細な手法と表現学習を行うために改良した手法について述べる．条件という用語が登場するが，ここでの条件とはサンプリングするときにどのような画像を生成するかを制御する対象を総称してそう呼ぶ．トレーニング時，デノイズアーキテクチャに条件を加算や乗算，結合などで与えながら学習することで，サンプリング時に条件に沿った画像を生成することが可能になる．

2.1 DiT を使用した画像生成

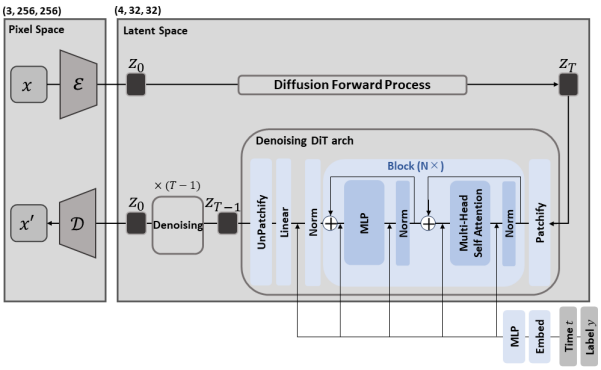

図 1: クラス条件付き潜在拡散モデル

図 1 は，DiT[13] のアーキテクチャを表している．VAE 潜在空間上での順拡散プロセスと，逆拡散プロセスをモデル化している．デノイズアーキテクチャが Transformer[20] を用いて構築されており，タイムステップ t とクラスラベル y をそれぞれ埋め込んだ後，訓練時に乗算と加算でデノイズアーキテクチャに渡されている．

2.1.1 Training

Algorithm 1 Training: PyTorch pseudo code

```
for x, y in loader:
  z_0 = E(x)

  t = torch.randint(0, 1000, (z_0.shape[0],))
  noise = torch.randn_like(z_0)
  z_t = sqrt(ᾱ_t)*z_0 + sqrt(1 − ᾱ_t)*noise

  output = model(z_t, t, y)

  loss = mean((noise - output)**2)

  optim.zero_grad()
  loss.backward()
  optim.step()
```

入力 x は $256 \times 256 \times 3$ の画像を表す．この画像をエンコーダ ε に通し，$32 \times 32 \times 4$ の潜在変数 z_0 に圧縮する．ここでのエンコーダ ε は，Stable Diffusion[6] によって公開されている VAE を使用している．これは ImageNet[21] で学習されたオリジナルのモデルをさらに Laion-Humans, LAION-Aesthetics[22] によってチューニングされている．

順拡散プロセスで用いる β は線形にスケジューリングしており，ノイズ付与の計算に使用される α，$\bar{\alpha}$ は式 (1)，(2) で表せる．

$$\alpha_t = 1 - \beta_t \qquad (1) \qquad\qquad \bar{\alpha_t} = \prod_{s=1}^{t} \alpha_s \qquad (2)$$

これらのパラメータを使用してタイムステップに応じたノイズが付与される．これは式 (3) で表される通り，潜在画像 z_0 と完全なガウシアンノイズ ϵ を，タイムステップ t における $\sqrt{\bar{\alpha_t}}$ と $\sqrt{1-\bar{\alpha_t}}$ の値に応じて足し合わされる．

$$z_t = \sqrt{\bar{\alpha_t}} z_0 + \sqrt{1-\bar{\alpha_t}}\epsilon \qquad (3)$$

モデルにはノイズ画像 z_t とタイムステップ t，クラスラベル y を入力として与える．訓練の目的は，式 (4) の $L(\theta)$ を最小化することである．

$$L(\theta) = -\log p_\theta(x_0|x_1) + \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \qquad (4)$$

q と p_θ はともにガウス分布を想定しており，KL ダイバージェンスは 2 つの分布の平均と共分散で評価することが可能である．式 (4) の第二項である KL ダイバージェンス項はモデルの出力 ϵ_θ を導入することで，式 (5) の L_{mse} で表される．

$$L_{mse} = \| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) \|^2 \qquad (5)$$

本実験では，Nichol と Dhariwal のアプローチ [23] に従い，ϵ_θ を L_{mse} で学習し，逆拡散プロセスでの共分散 Σ_θ を $L(\theta)$ をフルに使用して学習する．

2.1.2 Sampling

Algorithm 2 Sampling: PyTorch pseudo code

```
label = [0, 1]
n = len(label)
z_t = torch.randn((n, 4, 32, 32))

for t in [1000, 999, ..., 2, 1]:
  output = model(z_t, t, label)
  e = torch.randn_like(z_t)
  z_{t-1} = 1/sqrt(α_t) (z_t - (1-α_t)/sqrt(1-ᾱ_t) output) + σ_t·e
  z_t = z_{t-1}
samples = D(z_t)
```

サンプリングは，純粋なガウシアンノイズからクラスラベルという条件を付与しながらノイズを除去していく過程のことである．z_t から z_{t-1} のデノイズ（式 (6)）を，タイムステップ $t = 1000$ から $t = 1$ まで繰り返し，ノイズの除去された画像を生成する．

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta) + \sigma_t \cdot \epsilon \qquad (6)$$

デノイズされた潜在画像を訓練時に使用していた VAE エンコーダとセットで訓練された VAE デコーダを使ってピクセル空間の画像を得ることができる．

2.2 潜在拡散モデルでの表現学習

大規模データセットを使用して事前学習したモデルを，解きたいタスクに合わせた少量のデータセットを使用してチューニングし，高性能なモデルを得るという目的のもと，これを拡散モデルで構築しようと考えた．

事前学習ということを考慮して，アノテーションなしで利用できる無条件潜在拡散モデルと，表現条件付き潜在拡散モデルを設定した．これらを訓練し，後続タスクとして分類問題を解くことを考える．事前学習済みモデルの重みパラメータをクラス条件付き潜在拡散モデルへロードし，ファインチューニングを行い，そのモデルからのゼロショットクラス分類手法（セクション 3.1.2）を用いて分類結果を得る．

2.2.1 無条件潜在拡散モデル

図 2 は，クラス条件付き潜在拡散モデルからラベル y を入力しないよう改良した無条件潜在拡散モデルである．

Pre-training

事前学習として，潜在空間上で無条件拡散モデルをトレーニングする．目的関数はセクション 2.1.1 のクラスラベル条件付き潜在拡散モデルと同じであり，主にモデルの

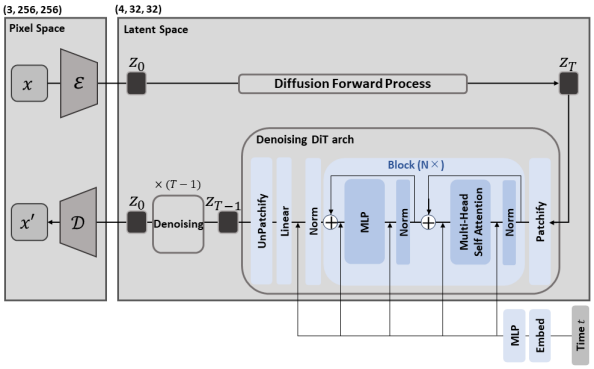

図 2: 無条件潜在拡散モデル

出力とガウスノイズとの平均二乗誤差を最小化するように学習される．

Fine-tuning

図 1 のクラスラベル条件付き潜在拡散モデルに対して無条件潜在拡散モデルの事前学習済み重みパラメータをロードし，クラスラベルを条件付けとして与えながら，ファインチューニングを行う．

2.2.2 表現条件付き潜在拡散モデル

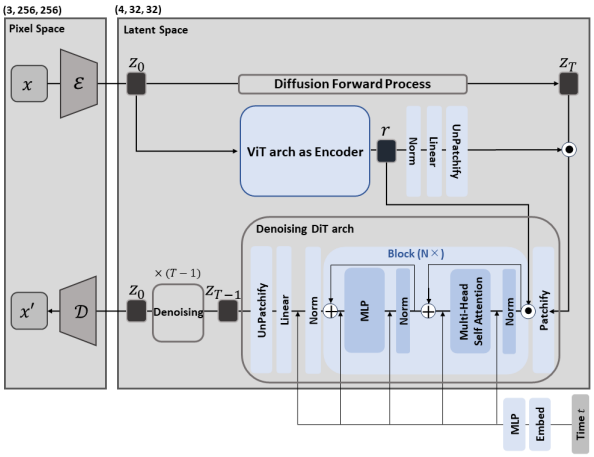

図 3: 表現条件付き潜在拡散モデル

図 3 は，拡散モデル内に ViT[24] を組合せ，明示的に表現を組み込んだ拡散モデルである．ViT の入力サイズは 32 × 32 × 4 であり，パッチサイズを 2，埋め込み次元を 768 に設定しているため，表現 r の次元は 256 × 768 である．その表現は，2 か所に渡される．1 つが，Norm 層，Linear 層，UnPatchfy 層を通り 32 × 32 × 4 次元の表現と，デノイズアーキテクチャに入力する 32 × 32 × 4 の画像を結合し，32 × 32 × 8 でデノイズアーキテクチャに渡される．もう 1 つが，256 × 768 の次元のまま，デノイズアーキテクチャに N 個（DiT B では N=12）存在するブロックの先頭に結合され，256 × 1536 の形になり，Linear 層を 1 つ設け 256 × 768 の次元でブロック内部に処理が進む．

VAE での潜在空間上で ViT にエンコードされた画像表現をデノイズ時に渡す，この処理が表現を条件付けるとい

うことであり，サンプリング時に生成したい画像を表現によって制御することが可能になる．

Pre-training

事前学習は，VAE 潜在空間上で，ViT でエンコードされた画像表現を与えながらトレーニングされる．ViT も同時に訓練されるため，デノイズ時に条件付けられる表現はトレーニングが進むにつれて拡散モデルの条件として最適化されていく．インスタンスベースの潜在拡散モデルと異なる点は，すでに訓練済みのエンコーダから得られた表現を条件付けるのではなく，拡散モデル内部でエンコーダも訓練され表現が最適化される点である．

Fine-tuning

クラスラベル条件付き潜在拡散モデル（図 1）に対して，表現条件付き潜在拡散モデルの事前学習済み重みパラメータをロードし，クラスラベルを条件付けとして与えながらファインチューニングを行う．事前学習時，ノイズ画像と表現を結合し，チャネルを 2 倍にしてデノイズモデルに入力していたため，ノイズ画像部分を受容するパラメータのみロードする．また，表現を結合していたブロック先頭部分に関しても，一つ設けた Linear 層を無視することでパラメータ数の異なるモデルへのロードに対処した．

3. 実験結果

3.1 データセット

本研究では，脳画像データセットを使用して実験を行った．画像データは CT スキャンのスライスであり，サイズは 512 × 512 ピクセルである．本研究においては，先行研究 [25] での前処理が行われたアノテーションデータセットを使用した．hypodensities, margin irregularity, blend sign, fluid levels の 4 つの血腫マーカーはそれぞれ独立で重複可能である (マルチラベル問題である)．本研究では，主に血腫の有無に着目しているため 4 つの血腫マーカーのいずれかが存在すれば血腫である，いずれも存在しなければ血腫ではないというというアノテーションを設定した．

3.1.1 分割

CT スキャンは 12 の施設から収集しており，そのうちの 2 から 12 を事前学習用のデータ．1 を後続タスク用のデータとして使用した．データ数は表 1 の通りである．また，後続タスク用のデータはさらに，訓練データ，検証データ，テストデータに 8:1:1 で分割した（表 2）．

表 1: 事前学習と後続タスクに使用する CT スキャンスライスのデータ数

purpose	quantity
pre-training (2-12)	10,313
downstream task (1)	1,781

表 2: 後続タスクに使用する CT スキャンスライスの用途別データ数

type of data	quantity
train data	1,424
validation data	178
test data	179
all	1,781

3.1.2 血腫分類

拡散モデルは内部の情報をノイズとして扱っているため，直接的な表現は存在せず，拡散モデルを表現を出力するエンコーダと捉えることは困難であると考えられる．そこで，クラス分類に関して Alexander et al.[26] が提案した Diffusion Classifier を参考にする．これは，クラスラベル条件付けの拡散モデルに対してすべてのラベルを条件付けとして渡し，出力されるノイズと本来のノイズの平均二乗誤差が最も小さいとき最適なクラスと判断するというものである．出力されたノイズ ϵ_θ とモデルに入力するノイズ画像 z_t からタイムステップ $t = 0$ の元画像を予測した z_0' を得ることが可能である（式 (7)）．本実験では，この予測した元画像と本来の画像との平均二乗誤差を測る．このイメージを図 4 で示す．

$$z_0' = \frac{1}{\sqrt{\bar{\alpha}_t}}(z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta) \tag{7}$$

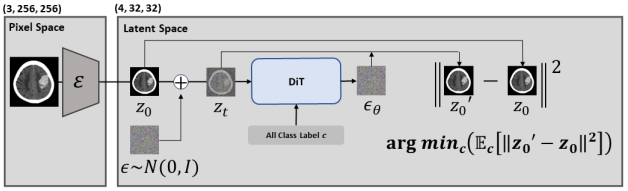

図 4: 拡散モデル分類器のモデル

この手法はどのタイムステップ t を選択するのが最適なのか明確に決まっていない，という欠点がある．我々が構築した拡散モデルは総タイムステップ数 $T = 1,000$ を想定しており，その時間に応じて拡散モデルが示す挙動は異なる．$t = 1,000$ に近ければ，99.9%以上がノイズ情報であり元画像の情報はほぼ失われていると思われ，画像の表現を学習しているとは思われない．$t = 200$ から 400 あたりの時間は，ノイズと画像どちらの情報もあり，これをデノイズすることを拡散モデルは学習するため，画像表現を強く学習していると思われる．現状，どの時間がクラス分類に最適かは貪欲に総当たりするしか方法がない．

3.1.3 生成モデルの評価方法

FID，sFID，IS(Inception Score)，Presicion，Recall の 5 つの指標で評価を行う．これらの値は，ImageNet で学習された Inception-v3[27] モデルを使用して計算される．

FID は，実際の画像と生成された画像の間の特徴距離を測定するための評価指標である．小さいほど良いモデルと考えられる．sFID(Spatial Frechet Inception Distance) は Nash ら [28] によって提案された FID の派生指標である．sFID は標準の’pool3:0’ からの特徴と，中間の’mixed_6/conv:0’ から 7 チャネルの特賞の両方を使用して FID を計算する．後者を含める理由は，モデル間の空間分布の類似性の感覚を提供するためである．IS(Inception Score) は，Inception-v3

による分類精度で画像の質を測り，すべての画像の分類結果数のエントロピーから多様性を測る．この値は高いほど，画像生成の質と多様性が保証される．ただ，Inception-v3 は ImageNet を使用して訓練されているため，今回実験で用いたような CT スライス画像を入力すると正しい結果が得られないと思われる．これらの指標で評価するにあたり，少なくとも 10,000 枚のデータ数が推奨されているため，我々の実験でも 10,000 のサンプルを生成して事前学習用のデータセットと比較する．

3.2 DiT を使用した脳画像生成

3.2.1 Training

事前学習用データを使用してクラス条件付潜在拡散モデルである DiT のトレーニングを行う．DiT[13] では，モデルの訓練ステップとして，training steps を導入しており，これは，データローダ 1 バッチの処理毎に加算される数値を表す．[*1]

本実験で使用したモデルは DiT B（Base）モデルであり，DiT[13] が提供しているモデルの大きさは S, B, L, XL の 4 種類ある．モデルの大きさは主に，埋め込み次元，ブロックの数，マルチヘッドアテンションのヘッド数の 3 つを制御している．詳細は表 3 に示す．B は埋め込み次元が 768，ブロック数が 12，ヘッド数が 12 である．画像のサイズは 256 にリサイズし，パッチサイズは 2，クラス数は 2 と設定して訓練した．

表 3: DiT モデルの詳細．[13] を参照．

Model	Blocks	Hidden size	Heads	Gflops
S	12	384	6	1.4
B	12	768	12	5.6
L	24	1024	16	19.7
XL	28	1152	16	29.1

3.2.2 Sampling

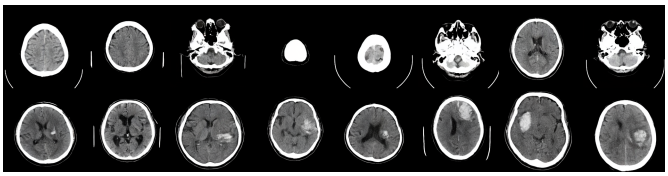

図 5: ラベル 0 で条件付けして生成した画像（上段）とラベル 1 で条件付けして生成した画像（下段）

図 5 の上段 8 枚は，クラスラベル 0 を条件として与えて生成した画像である．クラスラベル 0 は血腫がないことを意味しているため，ほとんどの画像において血腫は見当たらない．また，頭蓋骨が多くを占めるスライスには血腫が存在しないことが多いため，頭蓋骨のみの画像も生成されていることが確認できる．下段 8 枚は，クラスラベル 1 を条件として与えて生成した画像であり，頭蓋内に灰色のものが確認でき，血腫と判断できる．下段左から 2 枚目の画像は一見して血腫が確認できなかったり，一部は血腫の形が適切でなかったりなどうまく生成できない例もある．

3.3 潜在拡散モデルでの表現学習

3.3.1 Pre-training

事前学習用データを使用して，無条件潜在拡散モデルと表現条件付き潜在拡散モデルのトレーニングを行う．

それぞれの損失の推移は図 6 の通りである．この値は，

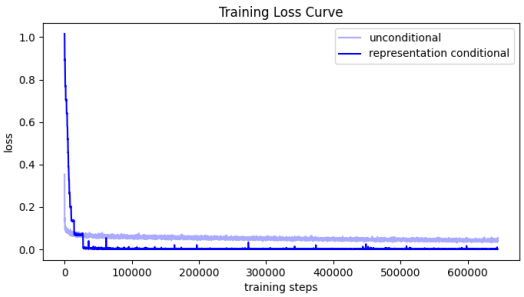

図 6: 無条件潜在拡散モデルと表現条件付き潜在拡散モデルのトレーニング損失の推移

式 (5) のノイズの平均二乗誤差と，式 (4) の $L(\theta)$ のフルの損失の和であり，後者はノイズ平均二乗誤差より 10^{-3} スケールほど小さく，主な訓練の目的はノイズ平均二乗誤差である．

無条件の方は，クラス条件付き潜在拡散モデルの損失の下がり方と類似している．しかし表現条件付きの方は，階段のようにある値を超えると急激に落ちるような下がり方をしている．この原因は不明だが，ほか二つと違う点は，モデル内に表現にエンコードする ViT が組み込まれており，それが損失を階段状にする要因ではないかと思われる．また，損失がかなり小さく下がっており，これは表現を条件付けて渡すことでモデルは表現から元画像情報を得ることができ，出力すべきノイズの予測が容易になっているためだと考えられる．

訓練におけるハイパーパラメータは，表 4 の通りである．使用しているデータ数とモデルの大きさ，GPU の容量などを考慮しバッチサイズを 16 に決定した．DiT[13] で行われた実験からは，トレーニングステップが増えるほど FID が小さくなるという結果が得られていた．本実験では，エポック数を 1,000 に設定してあるが，バッチサイズと考慮してもトレーニングステップ数は 600,000 ほどであり，この値は DiT[13] では十分な性能が得られていたためこの値

[*1] 例えば，今回使用したデータ数は 10,313，バッチサイズは 32 で設定したため，1 epoch あたり training step はおよそ 322（≒ 10,313/32）である．この実験では 1,400 epochs 回しているため，training steps はおよそ 450,800 である

表 4: 事前学習でのハイパーパラメータの値

epochs	1,000
batch size	16
model size	B
image size	256
patch size	2
learning rate	1e-4

を設定した．モデルサイズは実験の効率を考慮し B に設定した（モデルの詳細は表 3）．

3.3.2 Sampling

生成モデルの評価として表 5 を示す．比較のために，公開されている事前学習済みの VQ-GAN[29]，VQ-Diffusion[30] をファインチューニングして評価した結果も示す．元データと生成データの分布を比較したものが FID，sFID である．表現を条件付けて訓練させたモデルにおいて他と比べて小さい値であり，これは表現という元画像情報をサンプリング時に与えていることで元画像に似た画像を生成することができ，データ分布が必然的に近づいたためこの結果になったと考えられる．IS は生成画像の多様性や質を測る指標であるが，Inception-v3 は ImageNet で学習されているため，多様性という点で脳画像ばかり生成しているデータ分布は低く算出されていると考えられる．

表 5: 条件別潜在拡散モデルの評価結果と事前学習済み VQ モデルとの比較

Method	Conditions	FID	sFID	IS	Precision	Recall
VQ-GAN	None	27.69	43.10	2.981	1.00	1.00
VQ-Diffusion	None	22.55	52.95	2.795	0.98	0.54
DiT	Class Label	25.69	21.57	3.257	0.39	0.36
DiT	None	24.42	20.98	3.117	0.40	0.39
DiT	Representation	10.00	12.20	3.204	0.94	0.99

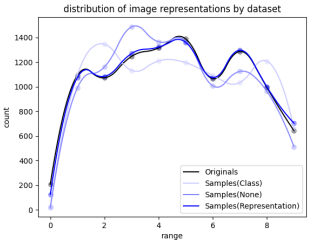

図 7: 事前学習用データの分布と条件別生成データの分布

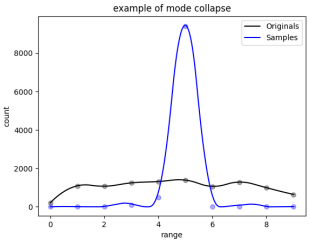

図 8: 事前学習用データの分布とモード崩壊時に想定される分布例

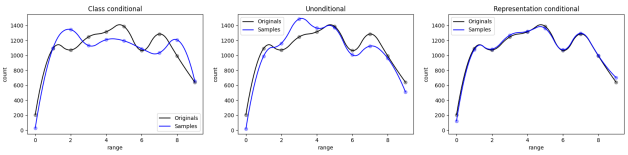

図 9: 事前学習用データと生成データの分布．(左) クラス条件付き，(中央) 無条件，(右) 表現条件付き．

図 7 は，画像データの分布を表している．横軸がデータが分布する範囲を表し，縦軸がその範囲 (離散値) に存在するデータ数を表す．Inception-v3 から出力された特徴表現 (2048 次元) を t-SNE[31] で 1 次元に圧縮，離散値に丸め込み，その個数をプロットした．図 8 は生成モデルがモード崩壊を起こした際に想定されるサンプル例を示したものである．モード崩壊とは，特に GAN における訓練で遭遇する問題であり，生成した画像の多くが，複製 (モード) を含み，変化に乏しい画像ばかりを生成してしまうことである．元データ分布では横軸の値が 5 の時におよそ 1400 枚で最もデータ数が多い．生成モデルは，それに似た画像を生成することばかりに囚われてしまい横軸 5 の値の画像ばかり生成してしまった例を示している．図 9 は図 7 を条件ごとに分けて表示した．本実験で訓練した潜在拡散モデルはどれもモード崩壊を起こしておらず，多様性のある生成ができていることが分かる．図 9 右の表現条件付きで生成された画像の分布は元データセットの分布と類似しており，FID の値を裏付ける分布になっていることが分かる．

3.3.3 Fine-tuning

事前学習で得られた無条件または表現条件付きの潜在拡散モデルの事前学習済み重みパラメータをクラス条件付き潜在拡散モデルにロードする．このモデルを後続タスクのデータを使用してクラスラベルでの条件を与えながらファインチューニングを行った．

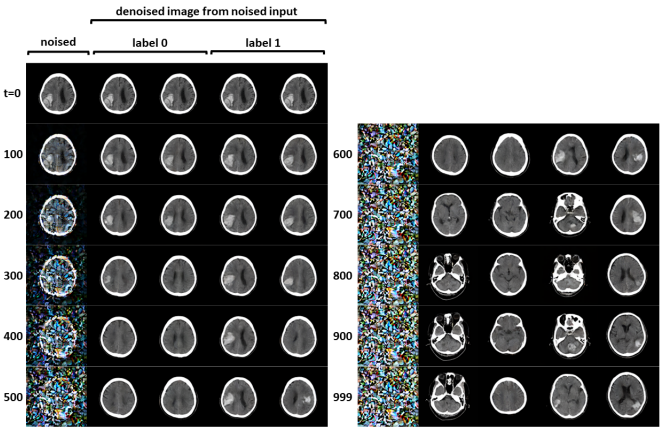

図 10: タイムステップ t におけるノイズ画像とそれを入力としてクラス条件を与えながら生成した画像例

図 10 は，ファインチューニングしたクラス条件付き潜在拡散モデルを使用し，タイムステップ t におけるノイズ画像からクラスラベルを条件付けながら t 回デノイズして得られた結果である．

例えば，$t = 500$ を見ると，元画像（$t = 0$ の一番左）に対して，$t = 500$ におけるノイズ画像が一番左の画像であり，骨の形が薄くわかる程度である．その画像を拡散モデルに入力し，右 4 枚のうち左 2 枚がラベル 0 を条件づけながら，右 2 枚がラベル 1 を条件づけながら，500 回デノイ

ズして得られた画像である．左 2 枚は元画像にあるはずの血腫が消えており，右 2 枚は血腫の位置が変わりつつも，血腫がわかる程度に生成されている．

この実験の意図することは，元画像の情報を壊しすぎずクラスラベルによる条件付けがうまく制御できるかを調べることである．タイムステップ t に応じて結果が異なることが分かり，$300 \leq t \leq 600$ では，元の脳の形を維持しながら血腫の有無を制御できている．

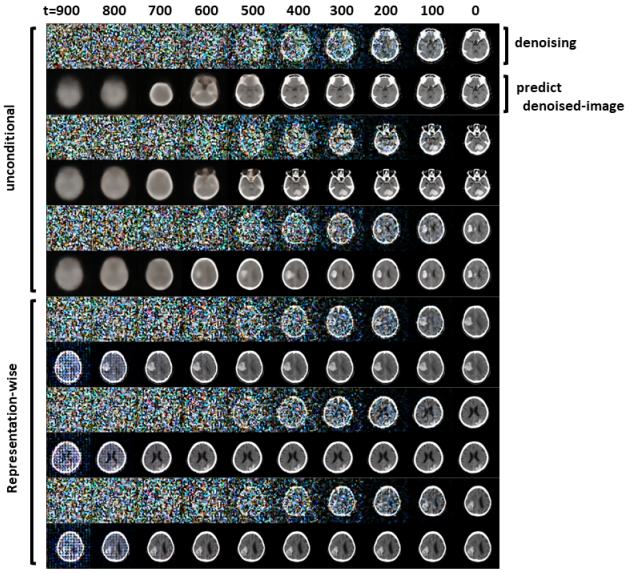

図 11: 無条件モデル（上 3 行）と表現条件付きモデル（下 3 行）における，純粋なノイズから画像をデノイズする過程（上段）とそのノイズからの元画像予測（下段）

図 11 は，上 3 つが無条件潜在拡散モデル，下 3 つが表現条件付き潜在拡散モデルでのサンプリングプロセスである．それぞれの上段が純粋なノイズから画像をサンプリングする過程，下段が上の画像からデノイズされた画像を予測したものである．無条件の方は $t > 600$ のノイズが多い画像からデノイズ後画像を正確に予測できておらず，ぼかしがかかっている．一方表現条件付きの方は，$t = 800$ でほとんどデノイズ後画像を予測できているように見て取れる．この元画像予測の違いが，無条件潜在拡散モデルと表現条件付き潜在拡散モデルの違いであると思われる．クラス条件付き潜在拡散モデルから分類器を抽出する際に使用するのはこの元画像を予測した画像であり，この予測画像と元画像との平均二乗誤差の大小で分類するため，元画像を予測する精度が高いほど分類精度も高くなると考えられる．

血腫分類を行う．任意のタイムステップ t から $t-1$ のデノイズを行う際に，データ数が多いほど，デノイズ時のランダム性に左右されず正確な結果が得られるため，本実験ではクラスラベル 0, 1 それぞれ 512 の計 1024 のバッチサイズでデノイズプロセスを行う．クラスラベル 0, 1 でそれぞれ平均二乗誤差は 512 個だけ得られ，その平均 2 つの大小を比較する．事前学習に無条件，表現条件付きで訓練された潜在拡散モデルのそれぞれの分類結果 (accuracy, F1 score) を図 12 で示す．

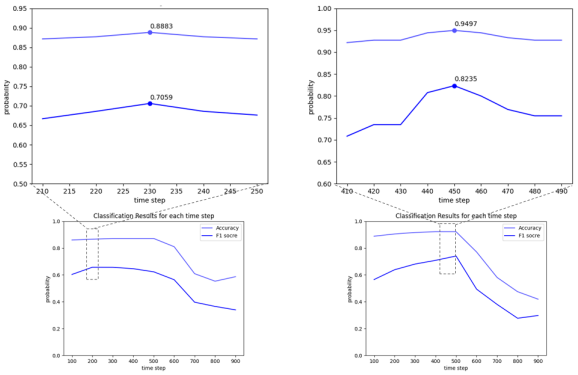

図 12: タイムステップごとの分類結果．事前学習：無条件（左），表現条件付き（右）．

図 12 の左は，事前学習時に無条件モデルを使用してファインチューニングしたモデルでのゼロショットクラス分類結果である．$t = 230$ で F 値が最大となった．また右は，事前学習時に表現条件付きモデルで訓練した場合の分類結果である．$t = 450$ で F 値が最大となった．まとめると，表 6 のようになる．

表 6: 後続タスクのテストデータでの血腫分類結果

Pre-train	Classification	Accuracy	F-score	Recall	Precision
None	ResNet50	0.7150	0.4950	**0.8928**	0.3424
None	ViT B	0.8659	0.5200	0.4642	0.5909
DINOv2	Linear	0.8882	0.6875	0.7857	0.6111
無条件 DiT	zero-shot	0.8882	0.7058	0.8571	0.6000
表現条件付き DiT	zero-shot	**0.9497**	**0.8235**	0.7500	**0.9130**

表現で条件付けして事前学習した潜在拡散モデルの分類結果が最も高い性能だった．我々が提案する潜在拡散モデルを用いた手法はどちらも，血腫分類において強力な対比学習手法である DINOv2 と比較して，accuracy では 7%の向上, F 値では 17%の向上を得られた．

4. 結論

本研究では，まず，Transformer を用いた拡散モデル (DiT) 内部に画像表現をエンコードする ViT アーキテクチャを組み込み，表現で条件付けられた画像生成モデルを提案した．このモデルで生成したサンプルはオリジナルの DiT と比較すると，FID が 15.69 向上することを示した．次に，自己教師あり表現学習として，拡散モデルを使用する手法を提案した．このモデルを用いた分類結果が DINOv2 を 7%上回ることを示した．さらなる今後の課題は，モデルサイズの調整，データセットの変更・拡大，また，条件付けなどで個人差を考慮した生成である．

参考文献

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[4] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, Vol. 22, No. 1, jan 2021.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 6840–6851. Curran Associates, Inc., 2020.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[7] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018.

[8] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[10] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[11] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about, 2022.

[12] Jeremias Traub. Representation learning with diffusion models, 2022.

[13] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[18] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021.

[19] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 25. Curran Associates, Inc., 2012.

[22] Romain Beaumont. img2dataset: Easily turn large sets of image urls to an image dataset. https://github.com/rom1504/img2dataset, 2021.

[23] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[25] Hokuto Hirano and Tsuyoshi Okita. Classification of hematoma: Joint learning of semantic segmentation and classification, 2021.

[26] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier, 2023.

[27] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[28] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W. Battaglia. Generating images with sparse representations, 2021.

[29] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021.

[30] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. *arXiv preprint arXiv:2111.14822*, 2021.

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, Vol. 9, pp. 2579–2605, 2008.