# 第6回定期ミーティング

2024/7/9

早稲田大学　基幹理工学研究科
電子物理システム学専攻 史研究室
石黒将太郎・野口颯汰

# Autodecoding Latent 3D Diffusion Models

Evangelos Ntavelis, Aliaksandr Siarohin, Kyle Olszewski, Chaoyang Wang, Luc Van Gool, Sergey Tulyakov

3Dオートエンコーダを用いて3Dデータセットから学習した特性を潜在空間に埋め込む
→一貫した見た目とジオメトリを持つ3D表現にデコードできる
→高品質な3Dモデルの生成が可能となる

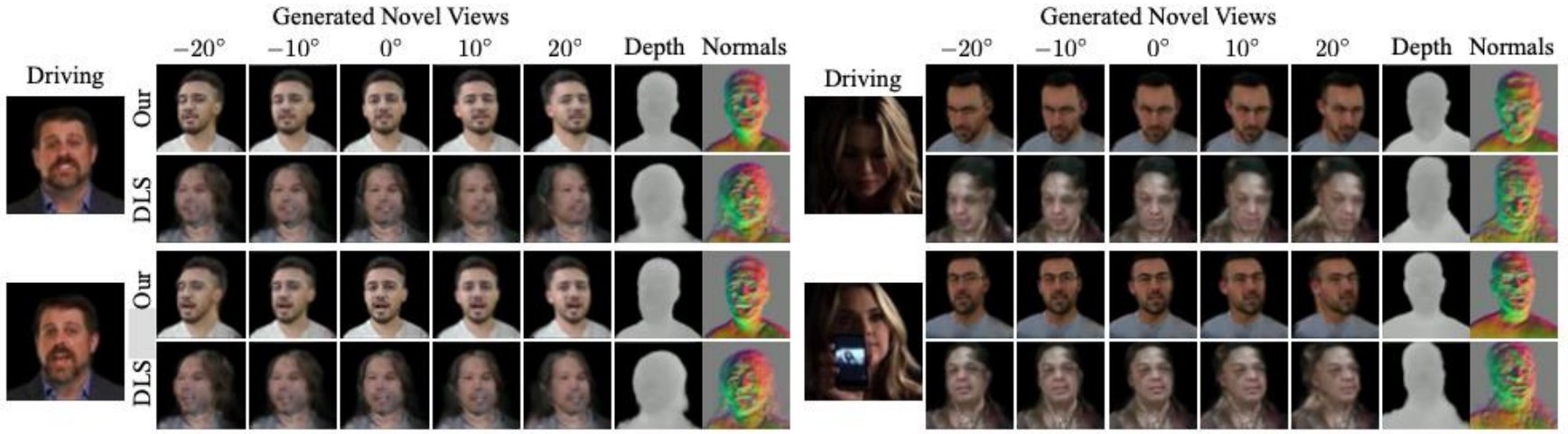


図1. 出力例[1]

背景：

- 現在利用可能なデータの多くは画像や単眼ビデオに限られており、一部の物体カテゴリーに対してのみ3Dメッシュやマルチビュー画像が存在しますが、これらは手間のかかるキャプチャプロセスを経て得られます。
- 拡散モデルのオートデコーダーを学習する為には大量のデータ必要だが、3Dデータセットが希少
  →椅子や机のデータセットに限定される
- 一次表現から最終的に放射および密度ボリュームに至るまでの過程で、各中間ボリューム表現が「ボトルネック」として機能する可能性があるという問題がある
- 2DではKLダイバージェンス正則化がボトルネックを解消するがオートデコーダーを使用した3D空間の学習プロセスでは別の正則化の方法を見つける必要がある

取り組んだ問題：

- さまざまな規模のデータセットで効率的に使用できる3D対応コンテンツのためのデノイジング拡散モデル
- 3D監督を必要としないオートデコーダーで潜在空間を学習する
- 事前に訓練された固定オートデコーダーのレイヤに適用できる堅牢な正規化および逆正規化操作を提案

概要

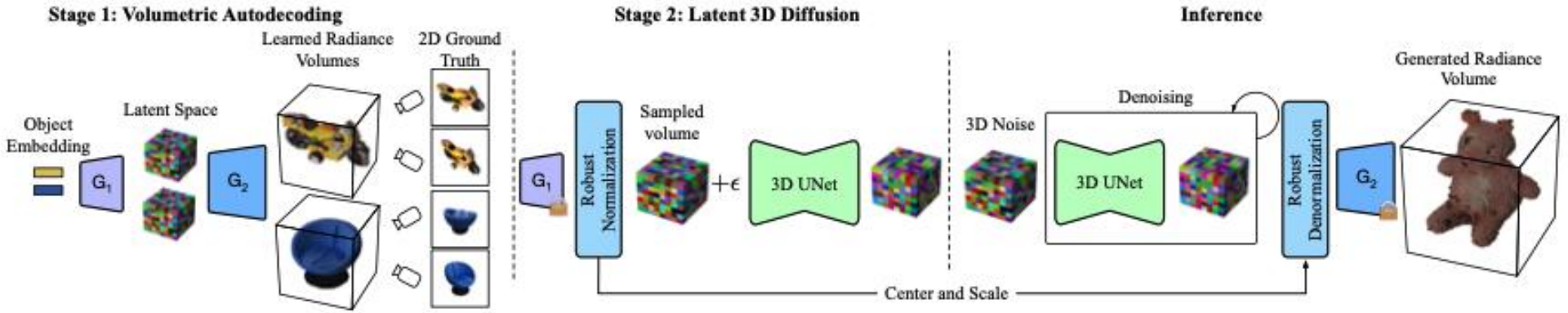


図1. 提案手法パイプライン[1]

Stage1 オートデコーダーの訓練
① $G_1$：各トレーニングデータセットオブジェクトに1次元の埋め込みを割り当てる
② $G_2$ ：放射フィールドに適した空間に変換

Stage2 拡散モデルの訓練
① $G_1$で潜在3D空間を生成
② 3Dデノイジング拡散プロセスを訓練

## 提案手法

**Canonical Representation（標準表現）：**

NeRFとは異なり、MLPを使用せず、デコードされたボクセルグリッドからトリリニア補間を使用して密度とRGB値を計算することで効率的かつ効果的な3D表現とレンダリングが可能

**ボクセルデコーダー：**

直接的に放射フィールドを表す密度と放射の3Dボクセルグリッドを生成することで
トレーニングや推論の際に大きな計算コストとメモリを必要としない

従来からの拡張

1. 埋め込みベクトルの長さ増加(64から1024)
   →より多くの情報を保持し、表現力を向上
2. 残差ブロックの増加(1から4)
   →深いニューラルネットワークの学習を容易にし、精度を向上させるため、これにより再構成品質が向上
3. Self-Attention層の導入
   →異なる部分間の関係を学習し、全体の一貫性を高める

## デコーダーの訓練

訓練方法：
デコーダーがレンダリングした画像とトレーニング画像との違いを最小化する

**Pyramidal Perceptual Loss** ：レンダリングされた画像とトレーニング画像の再構成ロスを計算
事前訓練されたVGG-19ネットワークの特定のレイヤーを使用して画像特徴を抽出し、
ピラミッドレベルごとに画像をダウンサンプリングして特徴間の違いを測定
→レンダリングされた画像がトレーニング画像に対して高い視覚的類似性を持つ

**Foreground Supervision** ：背景の影響を軽減する
レンダリングされた占有マップにL1ロスを適用して、それが画像に対応するマスクと
一致するように計算
→再構成されたオブジェクトの形状がより正確に維持され、視覚的な一貫性が向上

## デコーダーの訓練

非剛体オブジェクトの学習方法：

問題点

動的なポーズから被写体の形状や局所的な動きをモデル化し、対応する局所領域の非剛体変形も考慮する必要

解決策

小さな剛体成分のセットに分解し、それらのポーズを推定して標準3D空間で整列

各成分の変形を合理的に組み合わせるための手法

学習されたボリューム線形ブレンドスキニング（LBS）操作を行う
スキニングウェイトはトレーニング中に推定され、事前知識がなくても成分ごとに適切に割り当てられる
→各成分が合理的に変形し、全体として一貫性のある3D表現が可能

LDM

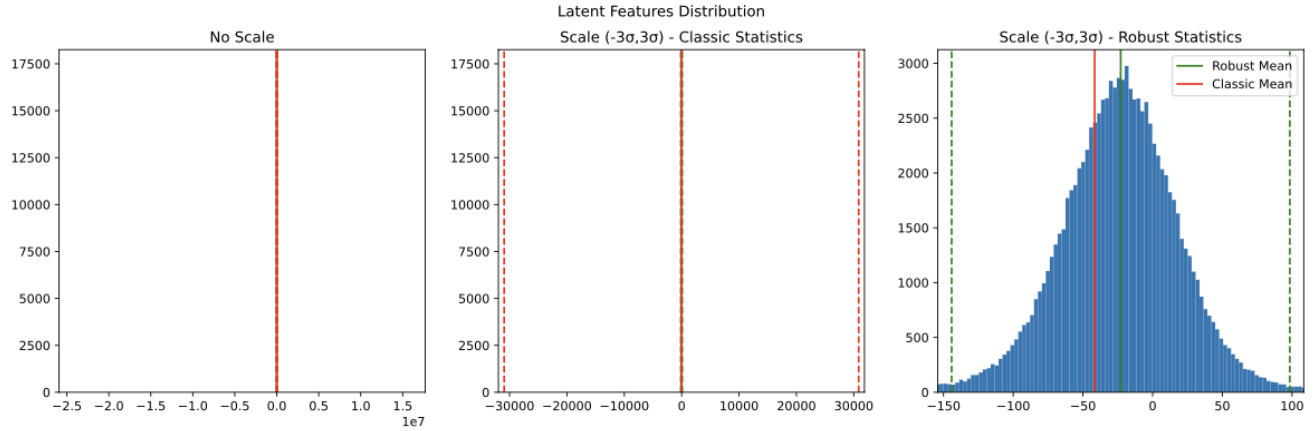


図2. 3Dオートデコーダーの潜在特徴分布[1]

**3Dオートデコーダーの潜在空間における特徴量の処理**

元々、3Dオートデコーダーの潜在空間における特徴量がベル型の分布を持っている
→事前分布を持たない潜在空間で操作することにより、可能なすべての潜在拡散解像度に対して単一のオートデコーダーを訓練
しかし、従来の方法とは平均と標準偏差が異なるので正規化を別途行う必要があり
→すべてのデータセットと訓練されたオートデコーダーに対して均一な拡散ハイパーパラメータを適用することが可能

サンプリング方法：
EDMを使用(2022時点のSOTA、複雑な仕組みを単純化し精度と速度を向上)

[1] Evangelos Ntavelis et al. Autodecoding Latent 3D Diffusion Models

## 無条件生成結果

- ✓ 幾何学とテクスチャの品質において実質的に高い忠実度を達成
- ✓ DiffRFと比べて高精度である

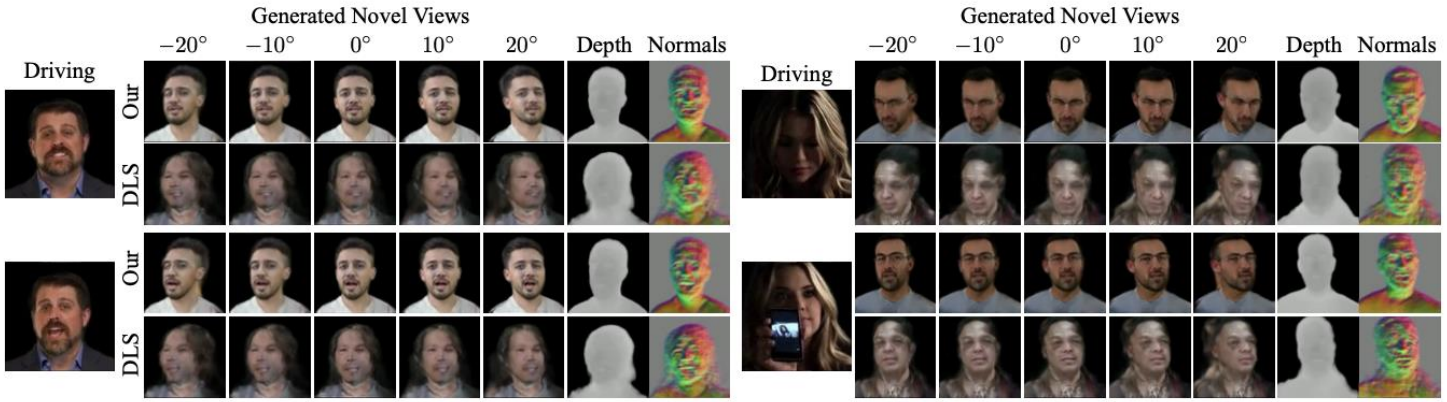


図3. CelebV におけるDirect Latent Sampling (DLS) との定性的比較 [1]

表1. 定性評価[1]

| Method | PhotoShape Chairs [57] FID ↓ | KID ↓ | ABO Tables [13] FID ↓ | KID ↓ |
|---|---|---|---|---|
| π-GAN [6] | 52.71 | 13.64 | 41.67 | 13.81 |
| EG3D [7] | 16.54 | 8.412 | 31.18 | 11.67 |
| DiffRF [49] | 15.95 | 7.935 | 27.06 | 10.03 |
| Ours | **11.28** | **4.714** | **18.44** | **6.854** |

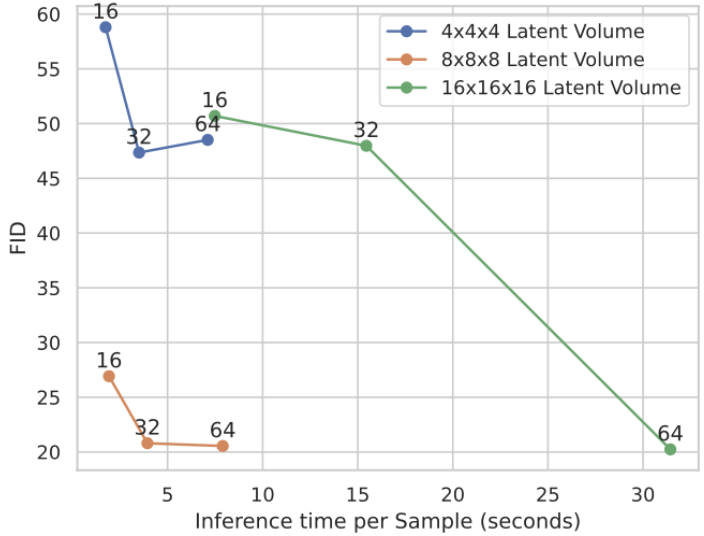


図4. サンプリングステップ・解像度と精度の関係[1]

## 条件生成結果

MiniGPT-4を用いてキャプションを生成
MiniGPT-4がしばしばオブジェクトの外観と一致しないテキストを生成するため、精度が低いこともあり

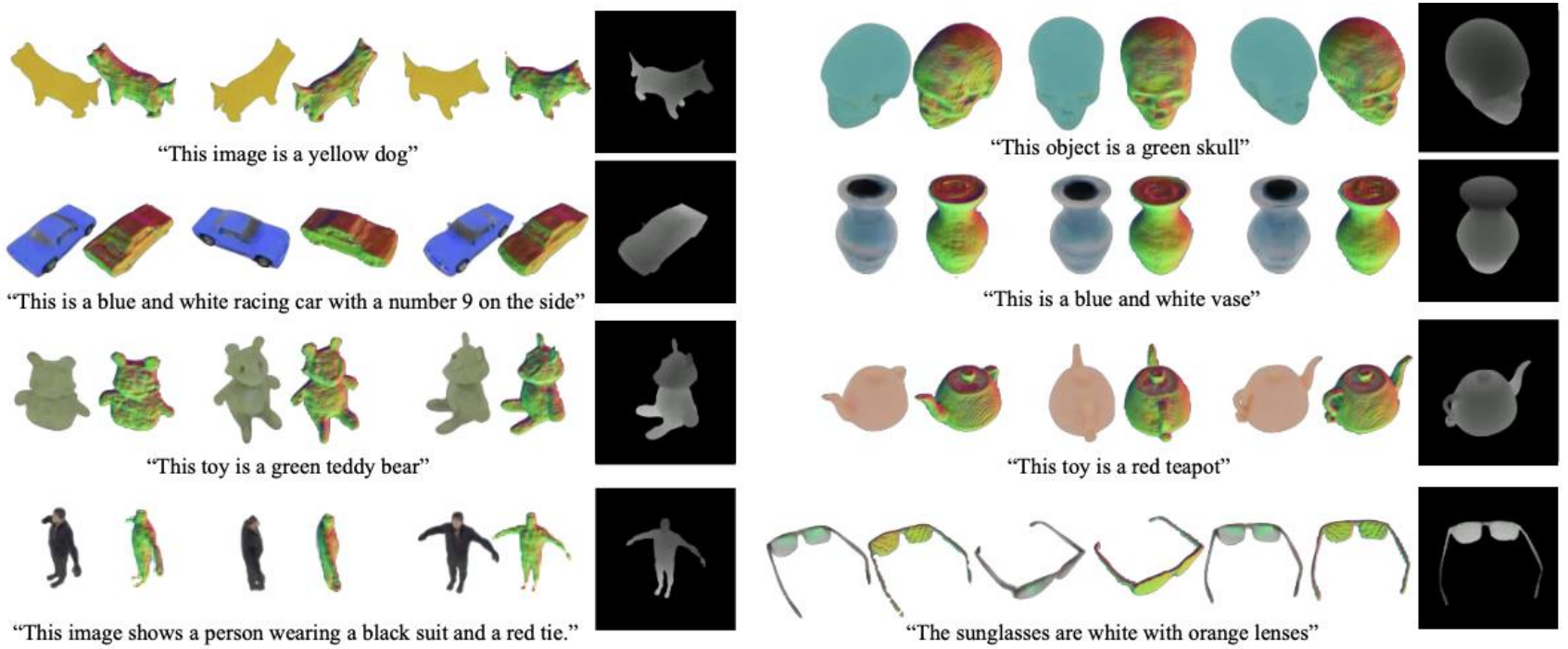


図5. Objaverseを用いたテキスト条件つき生成結果[1]

# 考察

- ✓ DiffRFと比べて潜在空間上で拡散モデルを動作させるので計算コストが少ないかつ物体の形とジオメトリを同時に生成することができるので一貫性が高い
  しかし、様々なオブジェクトに対応する必要はないかつ3Dデータセットは用意できる前提なので
  タスクに合わせたデコーダーを作成する必要あり

- ✓ 非剛体オブジェクトに対応するモデルであるため、無表情から笑顔になる動画データセットで追加学習することでより多様性がある生成が可能になる

- ✓ 独自データセットやタスクに特化したエンコーダー・デコーダーを作成する際に既存の3D-Unetを用いた拡散モデルと分散・平均値等で整合性をとる処理が必要である

[1] Evangelos Ntavelis et al. Autodecoding Latent 3D Diffusion Models

# 提案モデル

ー訓練

1. iPadのTrue-Depthセンサーを用いて作成した3Dデータセットを制作
2. 3Dデータを圧縮するエンコーダーの学習
3. 3D顔生成のノイズにする確率過程を学習
4. 3DMMをベクトルに変換するエンコーダーの学習

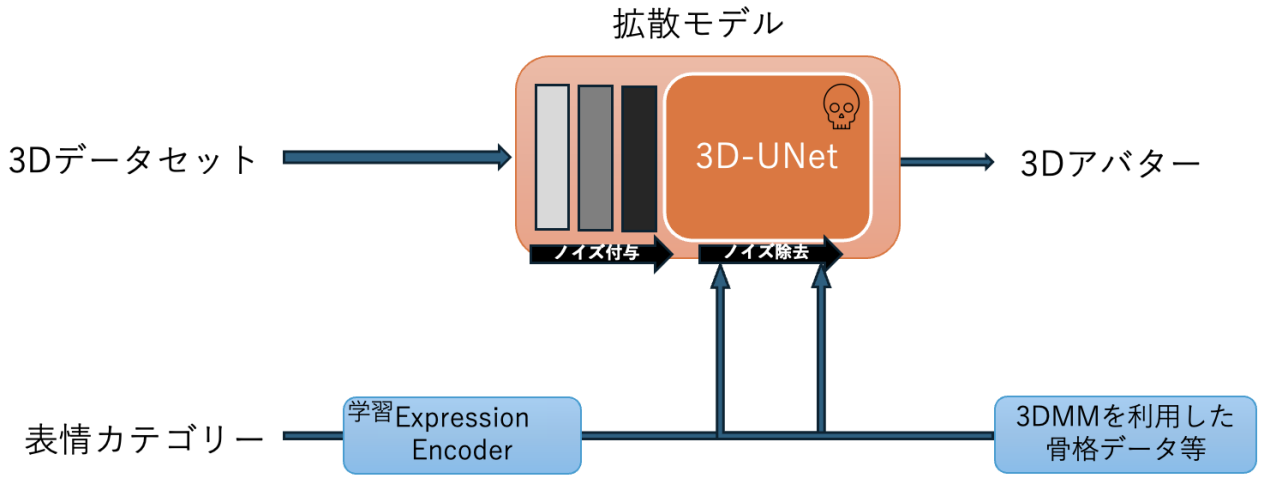


図. 提案モデル

ー推論

1. 表情変換したい人の無表情3Dデータを取得
2. 3D顔データを3DMMエンコーダーに入力
3. エンコーダーで得られたベクトルを3D U-Netにクロスアテンション層を用いて条件付け
4. 任意の表情に変換した3Dアバターの生成

# 研究計画

表. 今後の研究計画

| | 7月 | 8月 | 9月 | 10月 | 11月 | 12月 |
|---|---|---|---|---|---|---|
| データセット考察 | → | | | | | |
| 拡散モデル調査 | → | → | → | → | → | → |
| 実装と検証 | → | → | → | → | → | → |

実装内容

2D Stable Diffusionを0から学習 → 2D Stable Diffusionに表情条件付けエンコーダー搭載 → 3D U-Netを用いて3D Stable Diffusionを学習 → 3D U-Netを用いて3D Stable Diffusionを学習 → 3D Stable Diffusionに3DMM条件付けエンコーダー搭載

調査内容

- ✓ 生成結果の評価方法(3Dデータの生成精度、表情変換の生成精度)
- ✓ 3D U-Net文献調査、3D情報圧縮エンコーダー実現可能調査(少ないデータセットで可能なのか)
- ✓ 3Dデータ表現方法(NeRF、点群等)