第8回定期ミーティング

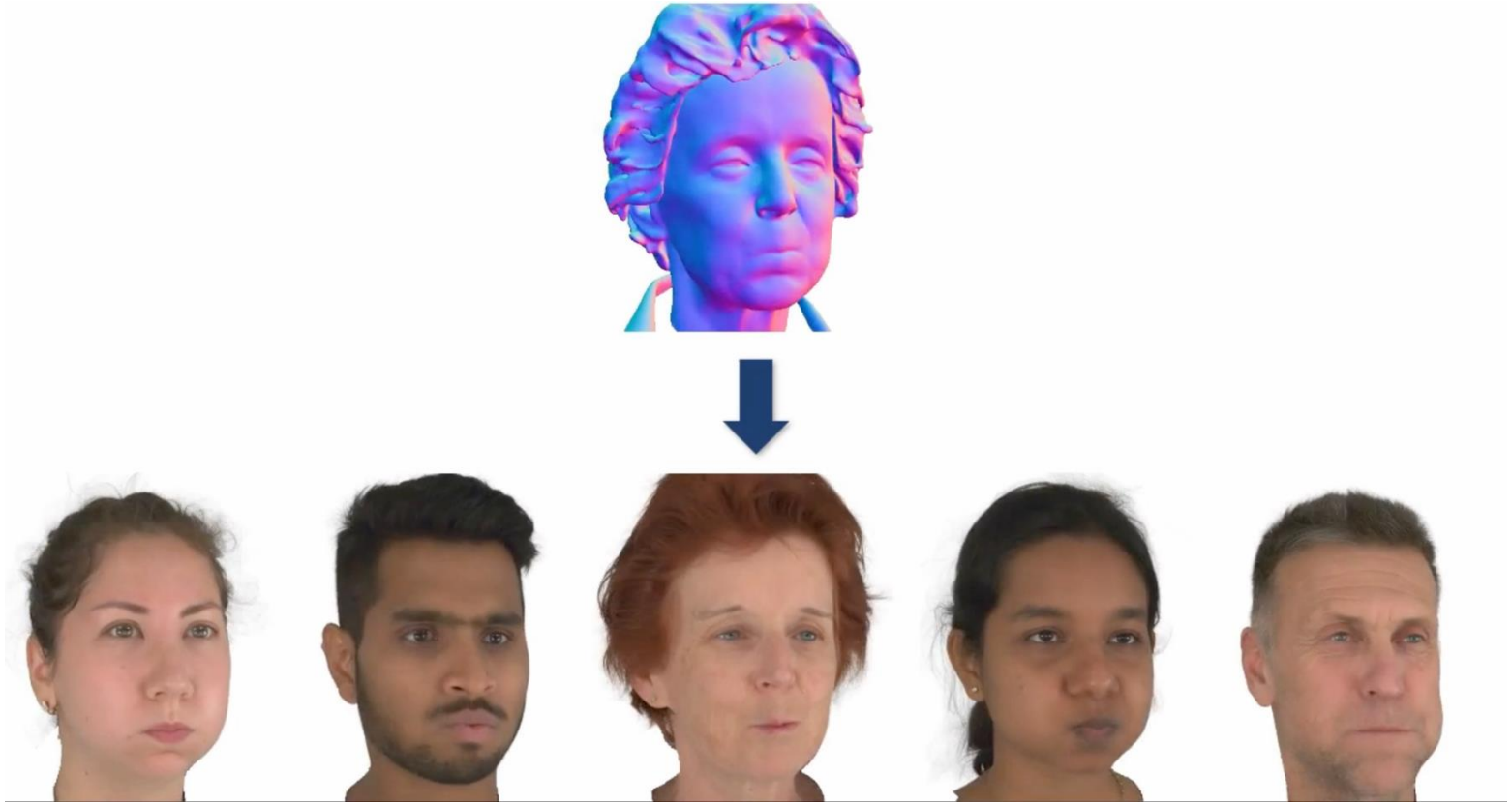
2024/10/22

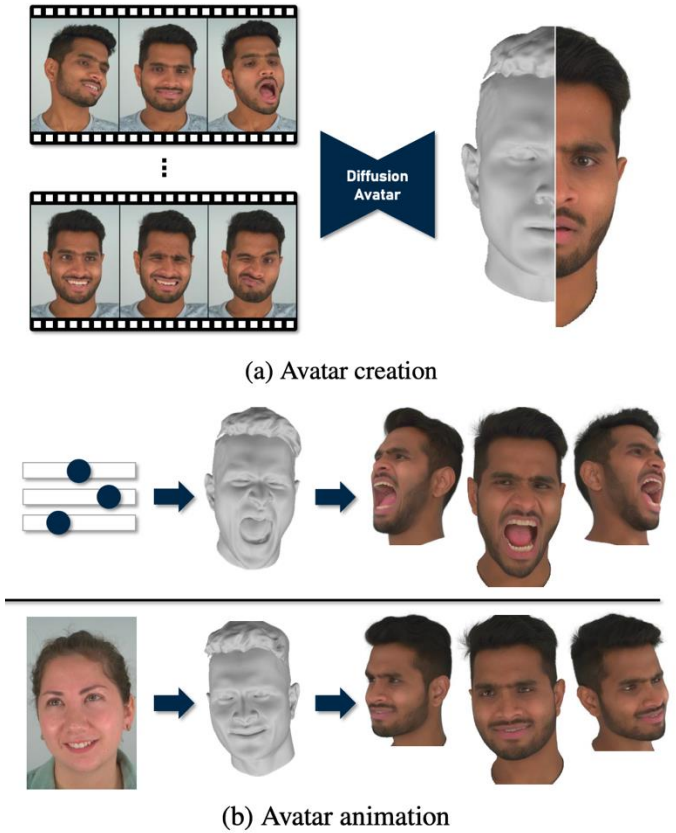
早稲田大学　基幹理工学研究科
電子物理システム学専攻 史研究室
石黒将太郎・野口颯汰

目次

DiffusionAvatars: Deferred Diffusion for High-fidelity 3D Head Avatars

Tobias Kirschstein Simon Giebenhain Matthias Nießner





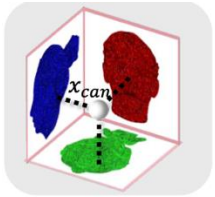
背景：

- 4Dフォトメトリック再構成問題
- 2Dニューラルネットワークや拡散モデルはフォトリアリスティックな画像生成に優れている一方で、3Dヘッドアバターのような動きや表情の自由な制御には限界がある
- 3D再構築手法は一貫したアニメーションが可能だが、レンダリングの品質が2Dモデルほど高くないことが課題

取り組んだ問題：

- 人物の追跡されたパラメトリックヘッドモデルを使用して、ポーズと表情を制御できるフォトリアリスティックな3D頭部アバターを作成
  - 拡散ベースのニューラルレンダラーに、空間特徴がリギングされたレンダリングされたNPHMメッシュを入力として使用
  - ControlNetを活用することで、未知の表情への汎化を促進
  - 拡散モデルの条件付けとして表情コードに追加で条件付けることで、複雑な表情を生成するモデルの能力がさらに向上

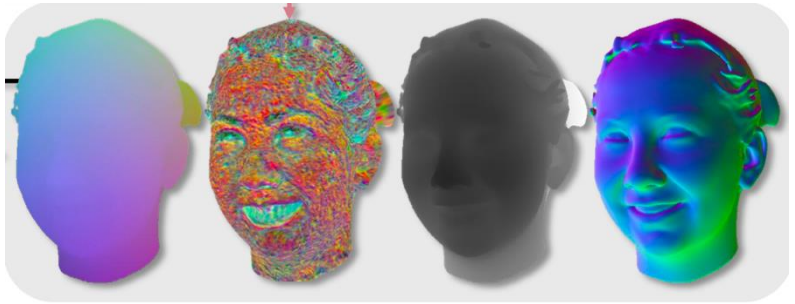
概要



TriPlane

NPHM

空間的特徴



<なぜTriPlaneが必要なのか？>

NPHMは陰関数表現のため、一貫したUV空間を持たない

そのため学習可能な特徴をメッシュ表面に結び付けるために使用される

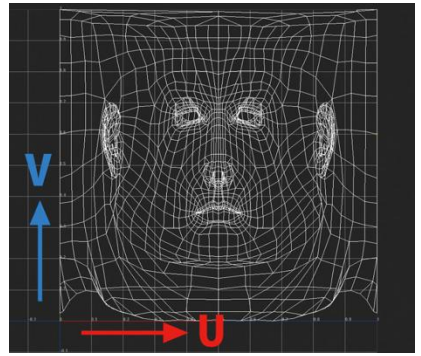
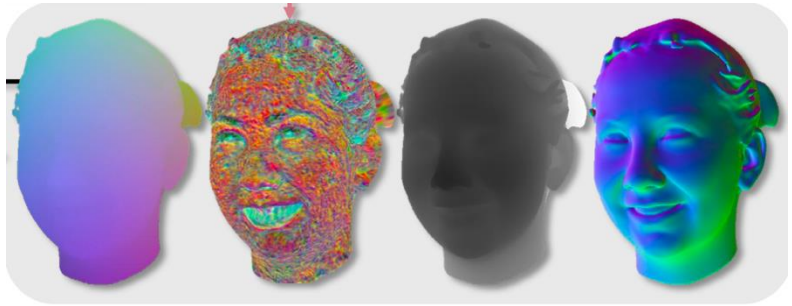


図2. UV空間例

概要



Control Net

深度・法線など

LDM(拡散モデル)

<なぜControl Netが必要なのか？>

LDMを、NPHMから生成されたラスタライズ画像を用いて条件付けするために使用される
高品質な画像の生成能力を維持しつつ、NPHMからの頭部形状とポーズの情報を利用できる

概要

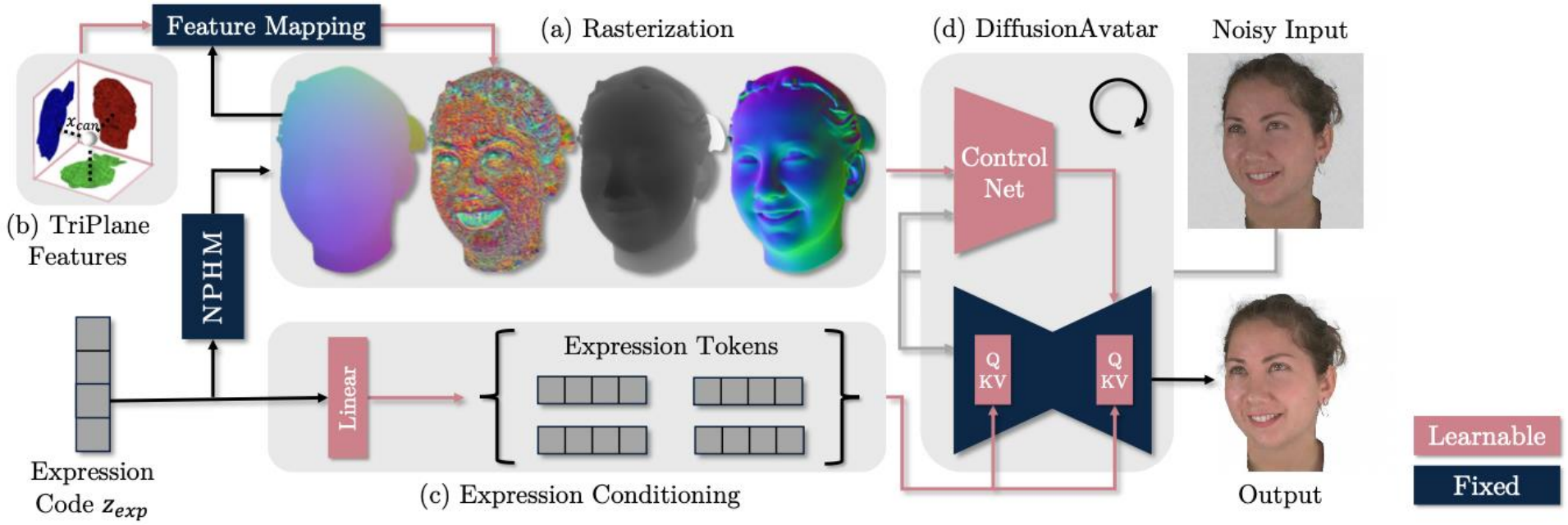


図. アーキテクチャ [1]

Deferred Diffusion：

暗黙的に定義されたプロキシ幾何形状の表面にリギングされた学習可能な特徴をデコードする拡散ベースのニューラルレンダラー

提案手法

拡散モデル：

v予測：ノイズ予測の代わりに、ノイズと元の画像の線形結合として新たな変数 v を予測する手法

メリット：

1.入力にすでに多くのノイズが含まれている場合でも、
損失がモデルに意味のある学習を常に導くため、収束が速くなる

2.純粋なノイズ入力でもモデルを訓練できるため、推論時に最も難しいデノイズステップに対応

$$v = \sqrt{\bar{\alpha}_\tau}\epsilon - \sqrt{1-\bar{\alpha}_\tau}x_0$$

提案手法

ラスタサイズ：

nvdiffrast(ラスタライザー)を用いて、拡散モデルに使用する実際の入力画像を生成
チャネルは法線、深度、NPHMメッシュの標準座標レンダリングで構成

TriPlane特徴マッピング：

学習可能な特徴を3D空間に結びつけ、ニューラルレンダリングで高詳細な画像生成を可能にする手法
NPHMがUVマッピングが存在しない為直接的なマッピングができない

なので、メッシュの表面に学習可能な特徴を結びつけるニューラルテクスチャを拡張するためには、、標準座標から空間構造を把握する必要がある
TriPlane：3D空間の各次元に学習可能な特徴を結びつける役割
AMBIENTMAP：NPHMメッシュの追加次元（口や目のトポロジーに対応）に特徴を結びつける

$$R^t_{\text{feat}} = \textsc{TriPlane}\left(R^t_{can,0\text{-}3}\right)$$
$$R^t_{\text{feat\_amb}} = \textsc{AmbientMap}\left(R^t_{can,3\text{-}5}\right)$$

提案手法

条件付け：

Control Netによる画像の条件付けは頭のポーズや大まかな表情を条件付けすることができる
しかし、詳細な表情はデコードするのが困難である

よって、IPAdapterに従い、新しいクロスアテンション層をU-Netに追加し、詳細な表情の合成を促進

損失関数：

$$\mathcal{L} = \mathbb{E}_{\epsilon,\tau,x_0^t,R^t,f_{exp}^t}\left[\left\|\mathcal{D}\left(x_\tau^t, R^t, f_{exp}^t\right) - v\right\|_2\right]$$

Control Net、表情条件付けのパラメータ、空間的特徴マップ TRIPLANE と AMBIENTMAPを訓練

トレーニング

データセット：NeRSemble(マルチビューデータセット)
→BackgroundMattingV2で背景削除し、セグメンテーションネットワークで胴体削除
→8人の動画(3300タイムステップ)に対して、NPHMをフィットさせ、メッシュを作成

訓練詳細：
Adam Optimizer
ControlNet・表情条件付け層の学習率：1e－4
TriPlane特徴検索の学習率：1e－2
バッチサイズ：8
解像度：H=W=512

→RTX A6000 GPUで約2日間

結果

表. 定量評価 [1]

Method	PSNR↑	LPIPS↓	JOD↑	AKD↓	AED↓	APD↓	CSIM↑
NeRFace [16]	23.0	0.279	6.76	5.37	1.06	0.053	0.787
DiffusionRig [12]	19.6	0.220	6.41	2.74	0.55	0.029	0.887
DNR [72]	24.5	0.226	7.32	2.06	0.63	0.027	0.903
DNR+GAN [72]	23.0	0.114	7.08	2.14	0.69	0.028	0.868
MVP [44]	23.6	0.221	7.02	3.42	0.78	0.034	0.882
Ours	**24.9**	**0.081**	**7.55**	**1.79**	**0.50**	**0.023**	**0.917**

図. 定性評価 [1]

考察

- ✓ 今回のモデルでは、自分でパラメーターを変更して表情を変更することが難しい
  顔の3Dモデルだけ今回のNPHMを用いることを検討
- ✓ しかし、それぞれのモデルが重すぎて、緩急的に困難

- ✓ Conteol Netの機構と3DMMを組み合わせることで軽量化と精度のバランスを保つ