第7回定期ミーティング

2024年7月22日(火)

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

アウトライン

- 論文紹介
  - HeadNeRF: A Real-time NeRF-based Parametric Head Model
- DiffusionRigに関する実験結果
- 参考文献

論文紹介

HeadNeRF：A Real-time NeRF-based Parametric Head Model

目的

- 優れたマルチビュー一貫性を持つNeRF構造を人間の頭部の表現に適用する
- IdentityやExpressionなどの意味的に分離された表現を実現

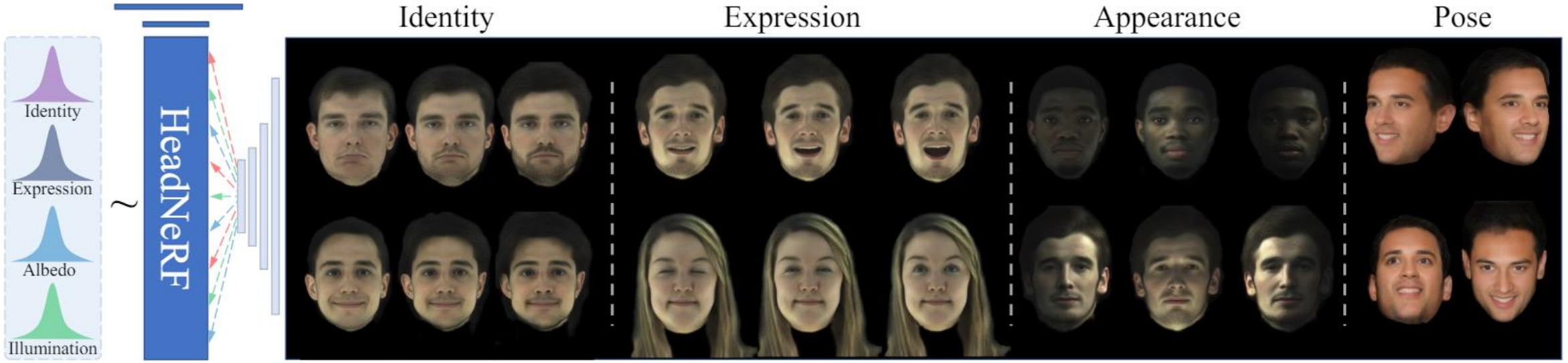


図. モデルのレンダリング結果[1]

[1] Yang Hong , ” HeadNeRF: A Real-time NeRF-based Parametric Head Model”, CVPR 2022, 2022-4-30

論文紹介

HeadNeRF：A Real-time NeRF-based Parametric Head Model

全体アーキテクチャ

- 3DMMと同様に、3D顔を制御するIdentity、Expressionとレンダリング結果を制御するalbedo、Illuminationをパラメータとして保持(初期値は3DMMと対応させることで効率的に学習)
- 4つのパラメータとNeRFで共通する位置エンコーディンベクトル$\gamma(x)$から$F(x) \in R^{256}$を予測
- $F(x)$のアップサンプリングとレンダリングを繰り返すことで最終的なレンダリング画像を取得

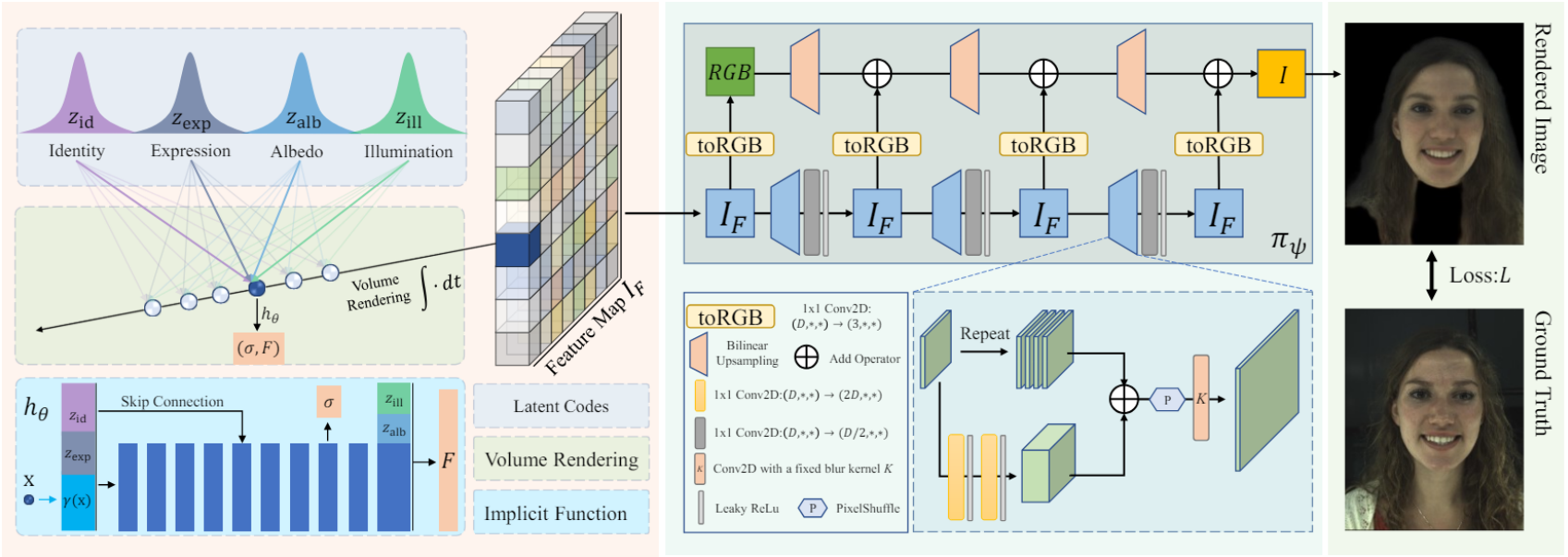


図. HeadNeRFの学習プロセス[1]

論文紹介

HeadNeRF：A Real-time NeRF-based Parametric Head Model

損失関数

- 頭部のみのレンダリング結果と実画像を比較

$$L_{\text{data}} = \left\| M_{\text{h}} \odot \left(\mathcal{R}(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{exp}}, \mathbf{z}_{\text{alb}}, \mathbf{z}_{\text{ill}}, P) - I_{\text{GT}} \right) \right\|^2,$$

- VGG16を用いて知覚損失

$$L_{\text{per}} = \sum_i \left\| \phi_i(\mathcal{R}(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{exp}}, \mathbf{z}_{\text{alb}}, \mathbf{z}_{\text{ill}}, P)) - \phi_i(I_{\text{GT}}) \right\|^2,$$

- 類似の表情や形状を持つ異なる被験者について、類似の潜在変数zを持つように学習が進むべき
  ⇨3DMMと対応した初期値と大きく離れないようにする(髪や歯などを扱う非表情属性は制約緩和)

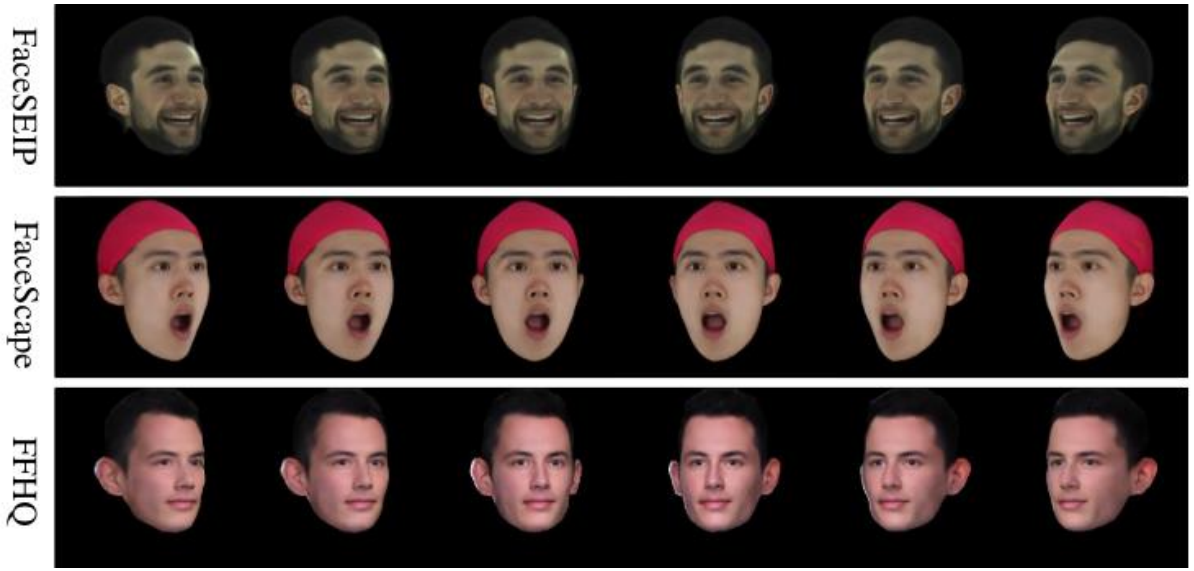
$$\begin{aligned} L_{\text{dis}} = & w_{\text{id}} \left\| \mathbf{z}_{\text{id}} - \mathbf{z}_{\text{id}}^0 \right\|^2 + w_{\text{exp}} \left\| \mathbf{z}_{\text{exp}} - \mathbf{z}_{\text{exp}}^0 \right\|^2 + \\ & w_{\text{alb}} \left\| \mathbf{z}_{\text{alb}} - \mathbf{z}_{\text{alb}}^0 \right\|^2 + w_{\text{ill}} \left\| \mathbf{z}_{\text{ill}} - \mathbf{z}_{\text{ill}}^0 \right\|^2, \end{aligned}$$

論文紹介

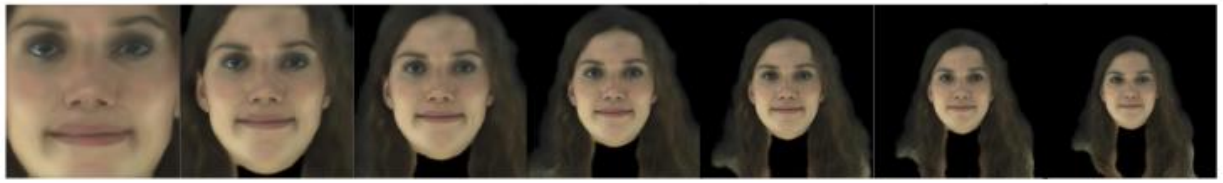
HeadNeRF：A Real-time NeRF-based Parametric Head Model

学習方法

- 学習パラメータ：$z_{id} \in R^{100}, z_{exp} \in R^{79}, z_{alb} \in R^{100}, z_{ill} \in R^{27}$
- ボリュームレンダリングの一部を2DNNレンダリングに置換することで推論を高速化
  ⇨40fps以上の速度で頭部画像をレンダリング可能
- 学習にNVIDIA 3090 GPU を3台使用(3日間訓練)



(a) Adjusting camera's position from near to far.



(b) Adjusting camera's FoV from small to large.

図. 2DNNレンダリングの妥当性とマルチビュー一貫性の保持[1]

論文紹介

HeadNeRF：A Real-time NeRF-based Parametric Head Model

レンダリング結果

- ある属性を表す潜在変数をサンプリングし、元の顔画像との補完画像をレンダリング

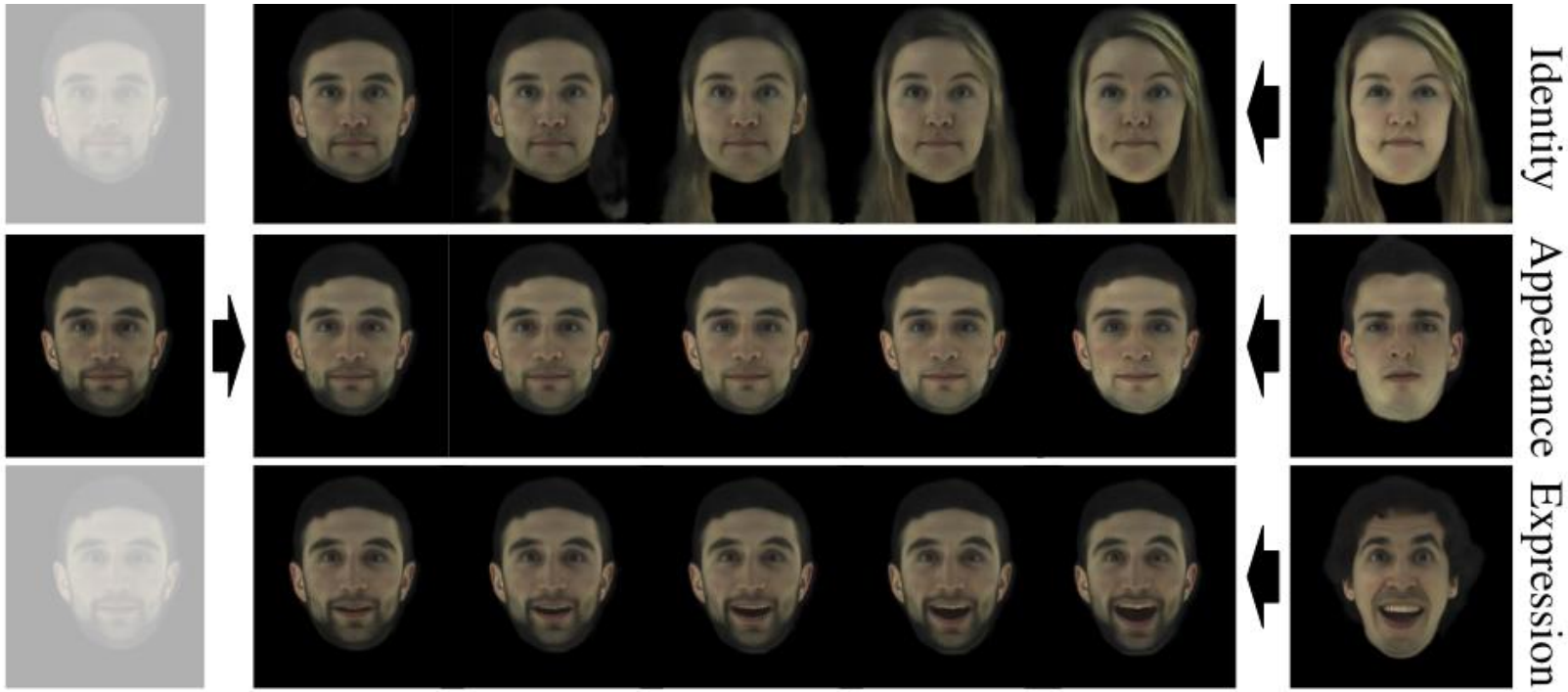


図. 潜在変数のサンプリングによる直接的な顔画像の編集[1]

論文紹介

HeadNeRF：A Real-time NeRF-based Parametric Head Model

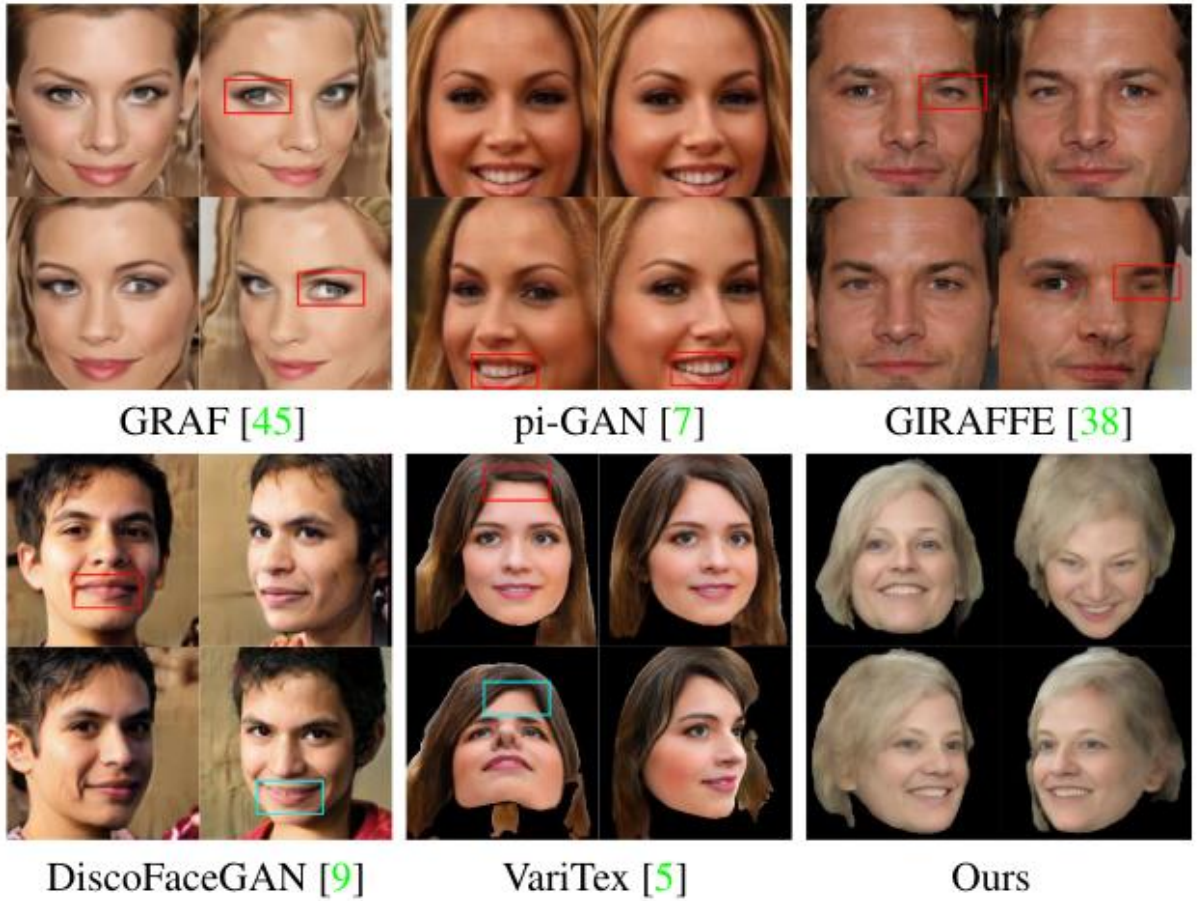
2DNNレンダリングの妥当性

- ボリュームレンダリングの一部を2DNNに置換することで、訓練期間が7日間から3日間に短縮、推論レンダリングは5秒から25ミリ秒に短縮、よりシャープな画像を取得することができる

図. 潜在変数のサンプリングによる直接的な顔画像の編集[1]

論文紹介

HeadNeRF：A Real-time NeRF-based Parametric Head Model



	CelebAMask-HQ			FFHQ		
Method	L1 ↓	PSNR ↑	SSIM ↑	L1 ↓	PSNR ↑	SSIM ↑
pi-GAN [7]	0.543	**24.8**	**0.799**	0.483	**24.9**	**0.810**
GIRAFFE [38]	0.420	20.6	0.628	0.428	20.3	0.635
Ours	**0.306**	19.6	0.702	**0.344**	21.6	0.755

図.他NeRF手法との比較[1]

DiffusionRigに関する実験結果

DiffusionRigの現状

- RTX 3080ti Laptopでの訓練では、顔の共通的な特徴を学ぶstage1に約1週間、個人アルバムから学ぶstage2に1時間必要
- 生成には1枚(256*256)で約20秒必要(DDIM20使用)
- 2D顔画像から3DMM(FLAME)を抽出するDECAモデルは容易に応用できそうだが、3DMM構築に必要なライブラリpytorch3dが曲者
- ライティングは転送しやすいが、表情を大きく変更させることは難しい

今後の課題(夏季休暇)

- 出力として3Dモデルを得ること、明示的な表情の変化(テキストベース?)、生成スピードの遅さを改善していきたい
- HeadNeRFの生成速度とパラメータ変化の妥当性は魅力的

図. 生成画像の例

HumanNorm: Learning Normal Diffusion Model for High-quality and Realistic 3D Human Generation

Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, Qing Wang

図. 出力例[1]

[1] N. Muller, Y. Siddiqui, L. Porzi, S. Rota Bulo, P. Kontschieder, and M. NieBner, “Diffrf: Rendering-guided 3d radiance field diffusion,” CVPR, 2023.

課題

テキスト→3D生成→リアルな人間生成×

原因：自然なRGB画像の生成特化→3D幾何構造や視点方向の認識が制限

解決策

法線適応型拡散モデルと法線整合型拡散モデルを学習

法線適応型拡散モデル： 3D形状の詳細を表す法線マップを生成
法線整合型拡散モデル：生成された法線マップに一致するリアルなカラー画像を生成

[1] N. Muller, Y. Siddiqui, L. Porzi, S. Rota Bulo, P. Kontschieder, and M. NieBner, “Diffrf: Rendering-guided 3d radiance field diffusion,” CVPR, 2023.

概要

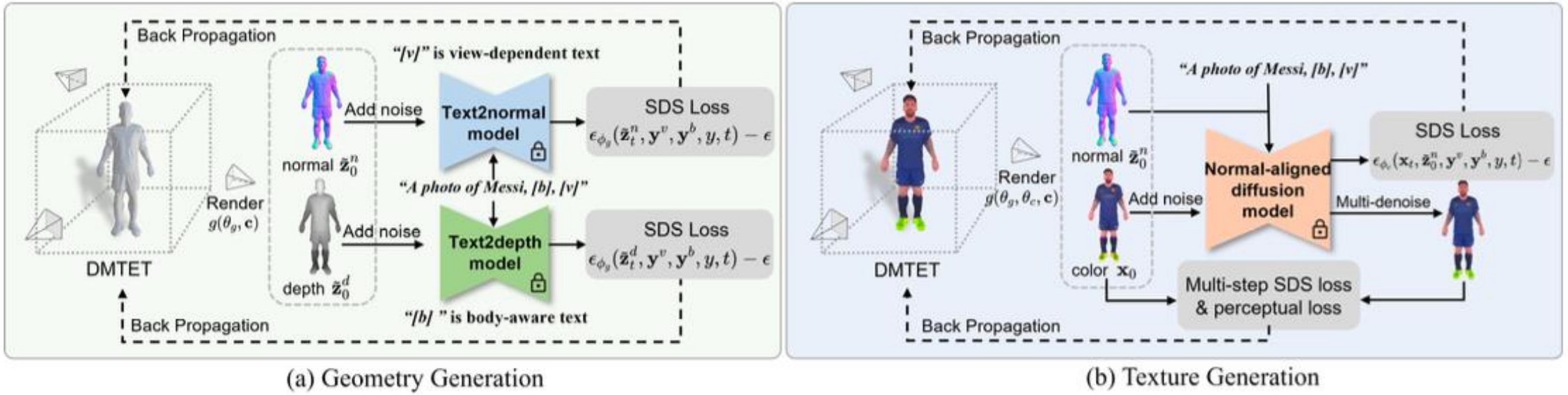


図. HumanNormパイプライン[1]

法線適応型
拡散モデル

目的：
複数の視点からレンダリングされた法線マップの分布を、理想的な法線マップの分布に一致

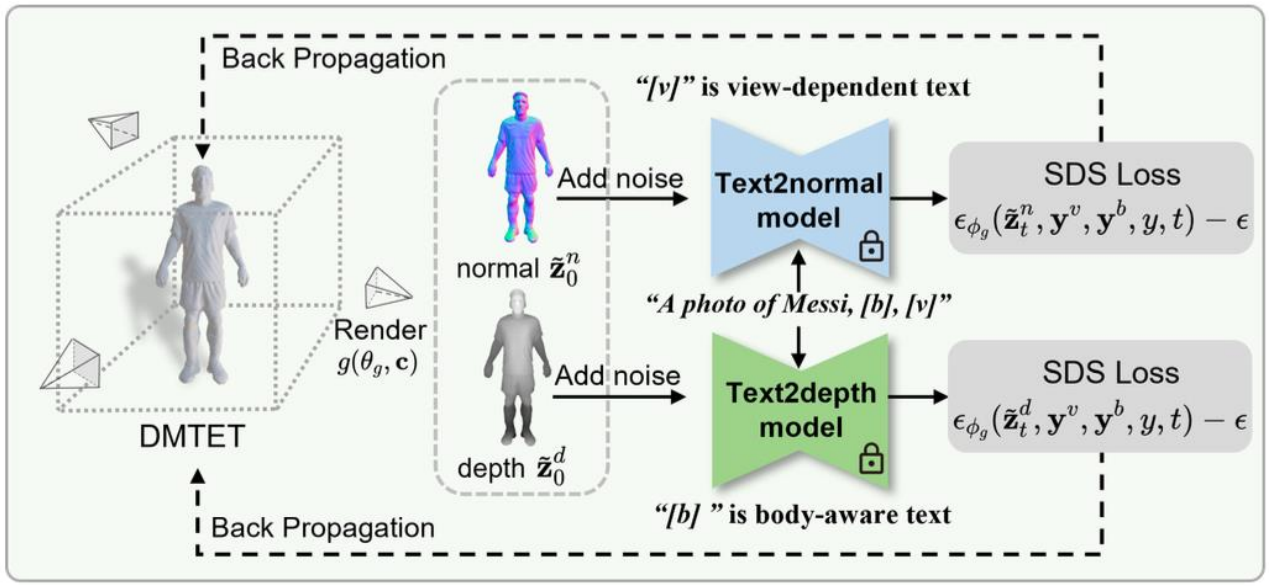


図. ジオメトリー生成[1]

3D人間データセットが乏しい
→Text2Image拡散モデルをText2法線マップ拡散モデルに対応させるためfine-tuning

レンダリングされた法線マップは視点角度の変化に伴い大きな変化→過学習または学習不足の問題が生じる可能性
→カメラパラメータの回転によって法線マップをワールド座標からカメラ座標に変換

変換した法線マップを訓練に使用

視点依存テキストyvと身体認識テキストybを追加の条件

[1] N. Muller, Y. Siddiqui, L. Porzi, S. Rota Bulo, P. Kontschieder, and M. NieBner, "Diffrf: Rendering-guided 3d radiance field diffusion," CVPR, 2023.

法線整合型
拡散モデル

目的：

生成されたテクスチャが幾何学的に
整合すること

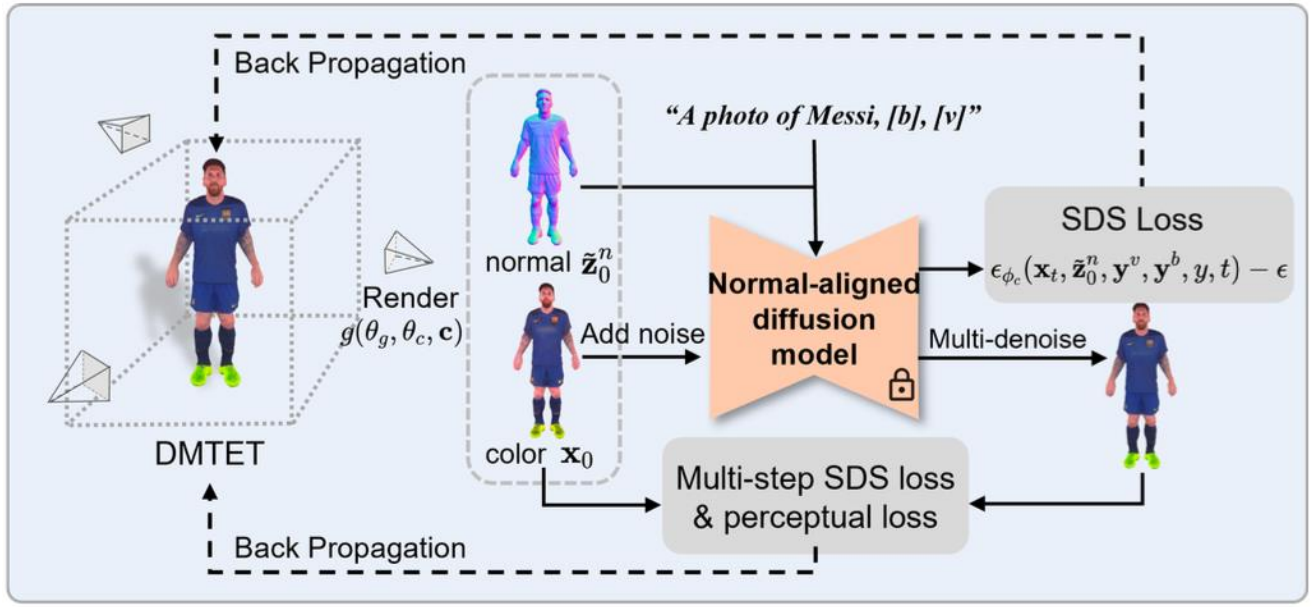


図. テクスチャ生成[1]

ControlNetを利用して、変換された法線マップをT2I拡散モデルのガイド条件として組み込む
→テクスチャ生成がより正確

[1] N. Muller, Y. Siddiqui, L. Porzi, S. Rota Bulo, P. Kontschieder, and M. NieBner, "Diffrf: Rendering-guided 3d radiance field diffusion," CVPR, 2023.

定性評価

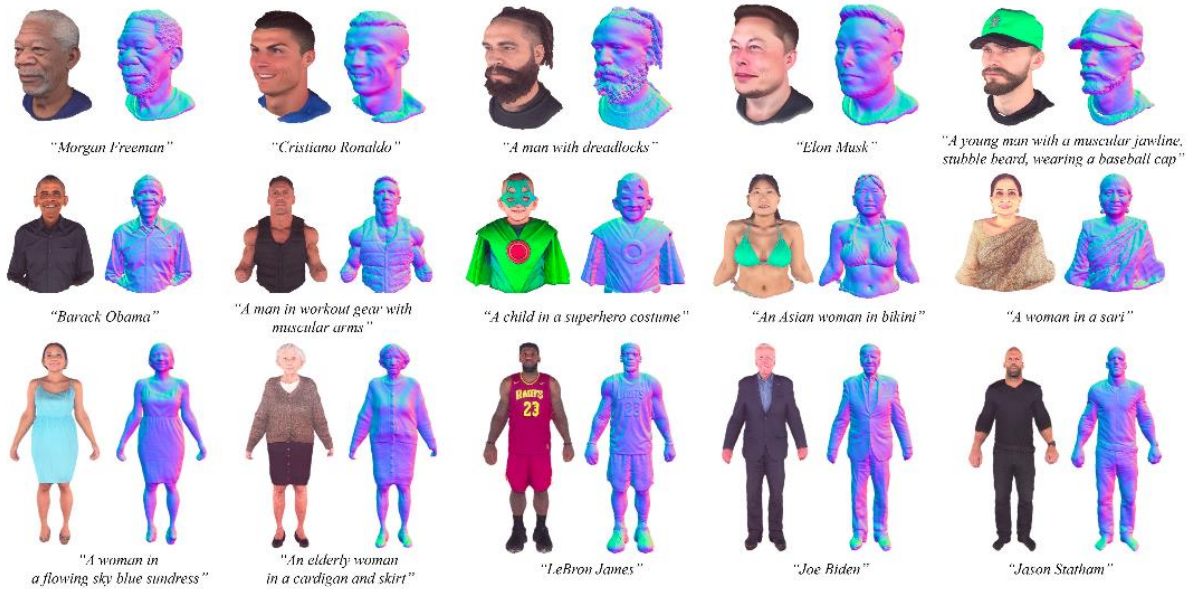


図. HumanNormによって生成された3D人間[1]

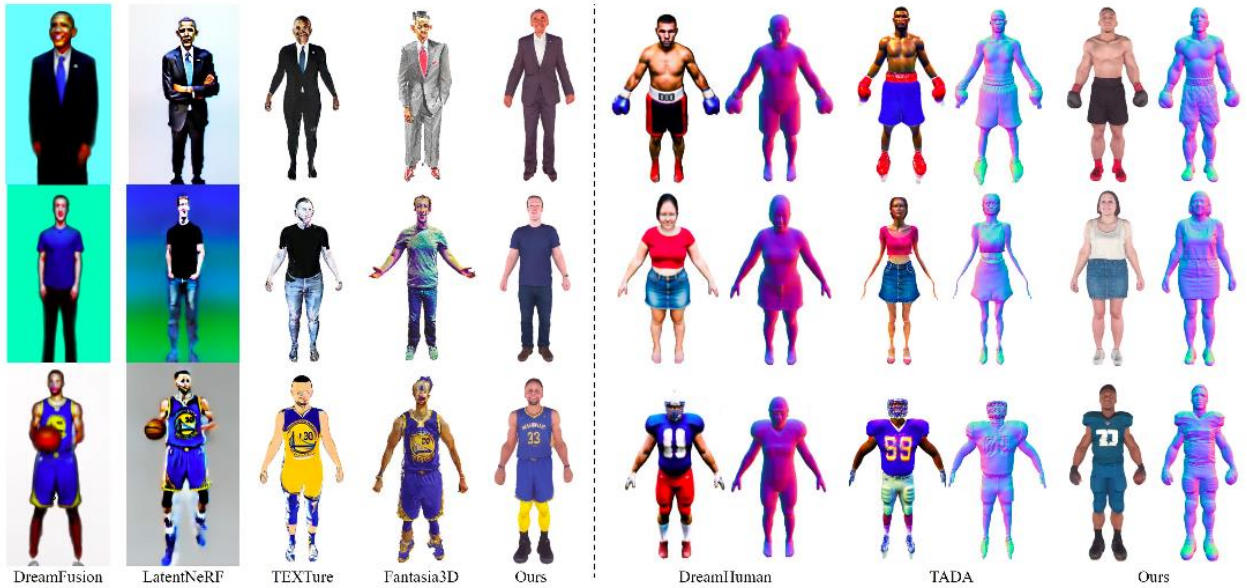


図. テキストから3Dコンテンツ生成比較[1]

[1] N. Muller, Y. Siddiqui, L. Porzi, S. Rota Bulo, P. Kontschieder, and M. NieBner, “Diffrf: Rendering-guided 3d radiance field diffusion,” CVPR, 2023.

定量評価

表. 定量評価

Method	FID ↓	CLIP Score ↑
DreamFusion	145.2	28.65
LatentNeRF	152.6	27.42
TEXTure	142.8	27.08
Fantasia3D	120.6	28.47
DreamHuman	111.3	30.15
TADA	120.0	30.65
HumanNorm (Ours)	92.5	31.70

生成された3D人間からレンダリングされたビューと、Stable Diffusion V1.5によって生成された画像の間のFIDを計算

CLIP
プロンプトと3D人間のレンダリングビューとの互換性

[1] N. Muller, Y. Siddiqui, L. Porzi, S. Rota Bulo, P. Kontschieder, and M. NieBner, “Diffrf: Rendering-guided 3d radiance field diffusion,” CVPR, 2023.

アブレーション評価

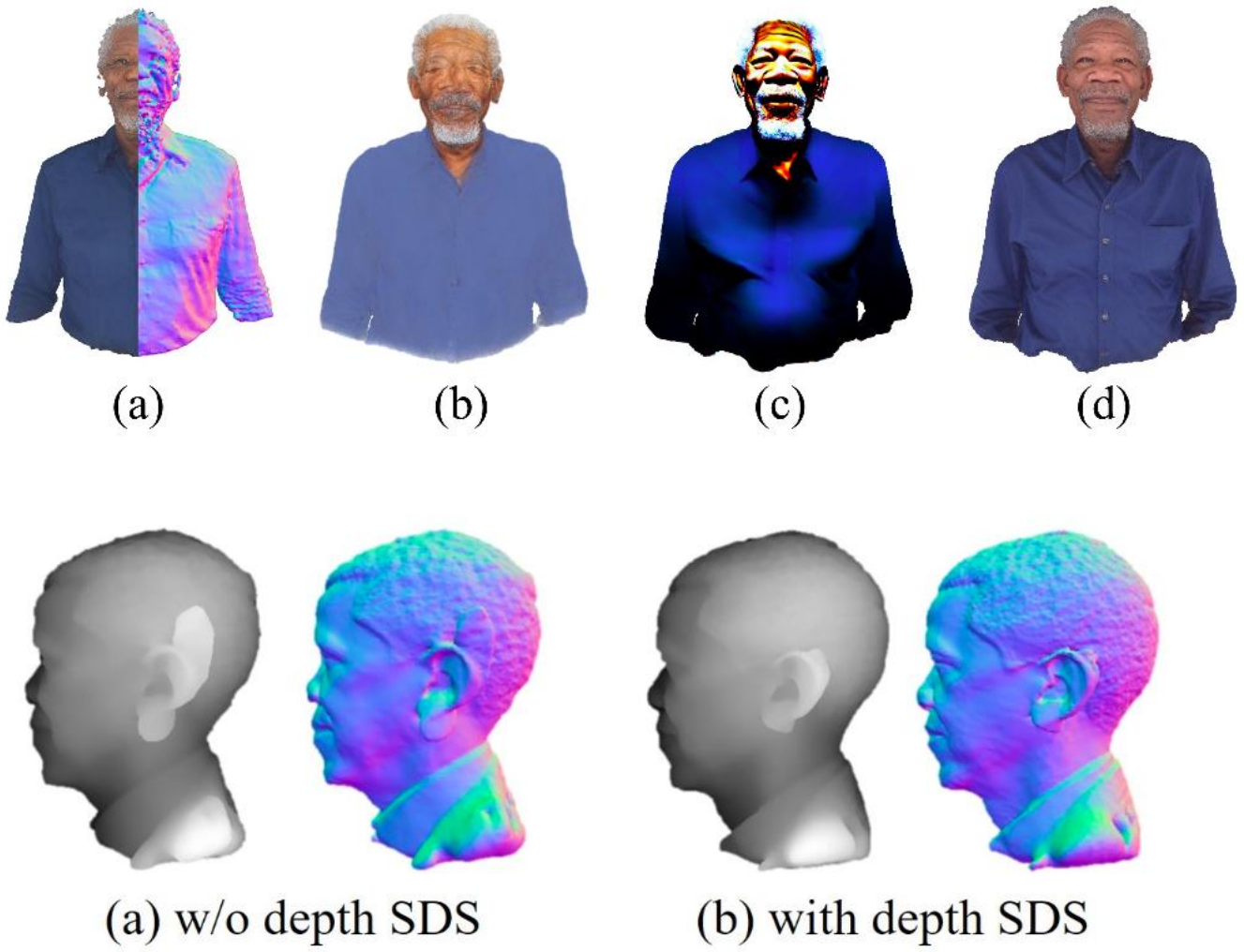


図. アブレーション研究[1]

[1] N. Muller, Y. Siddiqui, L. Porzi, S. Rota Bulo, P. Kontschieder, and M. NieBner, “Diffrf: Rendering-guided 3d radiance field diffusion,” CVPR, 2023.

今後の研究計画

表.今後の研究計画

	6月	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	→	→	→	→	→	→	→

Stable Diffusion・3DLDMコード解析

①：2Dto2D Stable Diffusionに表情カテゴリーエンコーダを組み込む
②：3DLDMを実装
③：3DLDMを３D顔画像で学習(ファインチューニング？)
④：表情カテゴリーを組み込んだ3D表情変換拡散モデル実装

深度情報欠如問題

- ✓ RGBD画像に対する畳み込み
- ✓ GCN
- ✓ VAEの調整
- ✓ 点群toRGBD変換の補完

3D生成技術

1.CLIP-Forge、DreamFields、CLIP-Mesh：
1. これらは、事前訓練されたCLIPモデルと3D表現を組み合わせ、CLIP損失を用いて3Dコンテンツを生成します。
2. **CLIPモデル**：テキストと画像を関連付けるために使用されるモデル。

2.DreamFusion：
1. SDS損失を導入し、テキストから画像への拡散モデルの監督下でNeRFを生成します。
2. Score Distillation Sampling（SDS）損失を通じて2D生成画像から3Dコンテンツを抽出

3.Magic3D：
1. NeRFとメッシュを使用して高解像度の3Dコンテンツを生成する二段階の方法を提案します。

4.Latent-NeRF：
1. 潜在拡散モデルを使用して潜在空間でNeRFを最適化し、画像のエンコードの負担を軽減します。

5.TEXTure：
1. テクスチャの生成、転送、および編集に特化した方法を導入します。

6.Fantasia3D：
1. 生成プロセスを幾何生成とテクスチャ生成に分解し、3D生成の性能を向上させます。

7.ProlificDreamer：
1. Variational Score Distillation（VSD）損失を提案し、高品質なNeRFを生成します。

8.IT3D：
1. GAN損失を導入し、生成された2D画像を活用して3Dコンテンツの品質を向上させます。

9.MVDream：
1. 一貫したマルチビューを生成するためのマルチビューディフュージョンモデルを提案しています。この方法は、複数の視点からの画像を統一的に生成しようとするものです。

10.Dream-Gaussian：
1. 3Dガウシアンスプラッティング技術を使用して、生成プロセスを高速化します。これは、3D点群を用いて効率的に3D構造を表現する手法です。

3Dアバター技術

- **EVA3D、LSV-GAN、GETAvatar、Get3DHuman**：これらはGANベースのフレームワークを用いて、3D人間生成のための3D表現を直接生成します。
- **AvatarCLIP**：CLIPを監督として使用し、SMPLとNeusを統合して3D人間を生成します。
- **DreamAvatar、AvatarCraft**：パラメトリックなSMPLモデルのポーズと形状を利用して人間の生成をガイドします。
- **DreamWaltz**：遮蔽対応のSDSおよびスケルトンコンディショニングを使用し、3D一貫性のある人間を生成します。
- **DreamHuman**：ポーズ対応のNeRFを使用してアニメーション可能な3D人間を生成します。
- **AvatarBooth**：顔と体を個別に扱う拡散モデルで、カジュアルな画像からパーソナライズされた人間を生成します。
- **AvatarVerse**：DensePoseを条件としてControlNetを訓練し、視点一貫性を向上させます。
- **TADA**：SMPL-Xを用いて階層的レンダリングとSDS損失を使用し、3D人間を生成します。

拡散ガイドによる3D生成フレームワークの基本的な考え方

フレームワークの目的：
•テキストyが生成対象として与えられた場合、このフレームワークの目的は、3D表現θからレンダリングされた画像x0を、2D拡散モデルの生成画像分布p(x0|y)に一致させることです。

画像のレンダリング方法：
•3D生成プロセス中に、レンダリングされた画像x0はランダムにサンプリングされたカメラcを使用して、微分可能なレンダリング関数g(θ, c)を通じて生成されます。
•これにより、さまざまな角度からの画像が生成されます。

最適化目標：
•レンダリングされた画像がさまざまな角度からの$q_\theta(x0|y) = \int q_\theta(x0|y, c)p(c)dc$として分布されると仮定します。
•フレームワークの最適化目標は、KLダイバージェンス（DKL）を最小化することです。具体的には、生成された3D画像分布$q_\theta(x0|y)$と、2D拡散モデルの生成画像分布p(x0|y)との間のKLダイバージェンスを最小化することです。

拡散ガイドによる3D生成フレームワークの基本的な考え方

最適化の難しさ：
•テキストから3Dコンテンツを生成する目的を直接最適化することは非常に困難です。
•このため、最近の方法ではSDS（Score Distillation Sampling）【31】やVSD（Variational Score Distillation）【48】といった損失を使用して、この最適化問題を解決しようとしています。

Fantasia3Dのアプローチ：
•Fantasia3Dは、3D表現θにおいて幾何学θgと外観θcを分離することを提案しています。
•この分離により、幾何学と外観を個別に最適化することが可能になります。

幾何学段階：
•幾何学段階では、レンダリングされた法線マップzn0の分布qθg(zn0 |y)を、自然画像分布p(x0|y)に一致させます。
•具体的には、KLダイバージェンス（DKL）を最小化することを目指します：min DKL(qθg(zn0 |y) ∥ p(x0|y))。

拡散ガイドによる3D生成の主なボトルネック

T2I拡散モデルの限界：
•T2I拡散モデルは、自然なRGB画像の確率分布をパラメータ化することに限定されています。
•このため、視点方向や3D幾何学を理解していません。

幾何学段階の問題：
•法線マップの分布を自然画像の分布に一致させる試みは、法線マップとRGB画像の性質が大きく異なるため、不適切です。
•この一致は幾何学の歪みやアーティファクトを引き起こします（図3(a)参照）。

テクスチャ段階の問題：
•外観分布と自然画像分布の間の発散を最小化することは、幾何学的なガイダンスが欠如しているため、偽の3D詳細を生じさせる可能性があります（図3(b)参照）。

appendix

法線適応型拡散モデルのfine-tuning最適化

$$\min_{\phi_g} \mathbb{E}_{\mathbf{c},t,\epsilon} \left[\| \epsilon_{\phi_g}(\alpha_t \tilde{\mathbf{z}}_0^n + \sigma_t, \mathbf{y}^v, \mathbf{y}^b, y, t) - \epsilon \|_2^2 \right]$$

cはカメラの姿勢、
tはタイムステップ、
εはノイズを表し、
yはプロンプトです。
σtとαtは拡散スケジューラーのパラメータです。
εϕg(·)は法線適応型拡散モデル

法線適応型拡散モデルのSDS損失

$$\nabla \mathcal{L}_{SDS}(\theta_g) = \mathbb{E}_{\mathbf{c},t,\epsilon} \left[\omega(t)(\epsilon_{\phi_g}(\tilde{\mathbf{z}}_t^n, \mathbf{y}^v, \mathbf{y}^b, y, t) - \epsilon) \frac{\partial g(\theta_g, \mathbf{c})}{\partial \theta_g} \right]$$

z˜ntは、タイムステップtでノイズεを含むレンダリングされた法線マップz˜0tに対応します。

ω(t)は拡散スケジューラーのパラメータであり、拡散プロセスの進行を制御します。

g(θg, c)は、幾何θgからカメラ姿勢cで法線マップをレンダリングすることを示します。

appendix

Progressive Geometry Generation

progressive positional encoding：
位置情報を段階的にエンコードし、生成プロセス中の詳細な位置情報の取り扱いを改善します。

詳細：
初期の最適化段階でノイズの多い表面を引き起こす可能性がある
→初期段階では、位置エンコーディングの高周波成分を抑制するためのマスクを使用
→ネットワークは低周波成分に集中し、トレーニングの安定性が向上
→トレーニングが進むにつれて、このマスクを徐々に減少させ、高周波成分を強調
→服のしわなどの細かい詳細を強化

progressive SDF損失：
SDF損失を段階的に適用し、幾何学的ノイズを軽減し、生成された幾何学の全体的な品質を向上させます。

$$\mathcal{L}_{SDF}(\theta_g) = \sum_{x \in P} \|\tilde{\mathbf{s}}_{\theta_g}(x) - \mathbf{s}(x)\|_2^2,$$

→ 複雑な形状を持つ人間モデルの生成において特に効果的

appendix

法線整合型拡散モデルの訓練目標

$$\min_{\phi_c} \mathbb{E}_{\mathbf{c},t,\epsilon}\left[\|\epsilon_{\phi_c}(\alpha_t \mathbf{x}_0 + \sigma_t, \tilde{\mathbf{z}}_0^n, \mathbf{y}^v, \mathbf{y}^b, y, t) - \epsilon\|_2^2\right]$$

マルチステップSDS損失

①法線整合型拡散モデルの基本的なSDS損失

$$\nabla \mathcal{L}_{SDS}(\theta_c) = \mathbb{E}_{\mathbf{c},t,\epsilon}\left[\omega(t)(\epsilon_{\phi_c}(\mathbf{x}_t, \tilde{\mathbf{z}}_0^n, \mathbf{y}^v, \mathbf{y}^b, y, t) - \epsilon)\frac{\partial g(\theta_c, \mathbf{c})}{\partial \theta_c}\right]$$

②マルチステップSDS損失

複数の拡散ステップを使用してRGB画像の分布を回復

→最適化中の安定性が促進され、局所最適解に陥るのを防ぐ

$$\nabla \mathcal{L}_{MSDS}(\theta_c) \approx \mathbb{E}_{\mathbf{c},t,\epsilon}\left[\omega(t)(h(\mathbf{x}_t, \tilde{\mathbf{z}}_0^n, \mathbf{y}^v, \mathbf{y}^b, y, t) - \mathbf{x}_0)\frac{\partial g(\theta_c, \mathbf{c})}{\partial \theta}\right] + \lambda_p \mathbb{E}_{\mathbf{c},t,\epsilon}\left[\left(V(h(\mathbf{x}_t, \tilde{\mathbf{z}}_0^n, \mathbf{y}^v, \mathbf{y}^b, y, t)) - V(\mathbf{x}_0)\right)\frac{\partial V(\mathbf{x}_0)}{\partial \mathbf{x}_0}\frac{\partial g(\theta_c, \mathbf{c})}{\partial \theta_c}\right],$$

③知覚損失

過度な飽和効果を防ぐ

レンダリング画像の自然なスタイルが法線整合型拡散モデルによって生成された画像と一致することを目指す

appendix

実装の詳細

幾何生成に15,000回の反復
テクスチャ生成に10,000回の反復

24GBメモリを搭載した単一のNVIDIA RTX 3090 GPU：生成に約2時間

最終的なレンダリング画像と動画の解像度は1024 × 1024