第４回定期ミーティング

2024年6月11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
  - DiffusionRig: Learning Personalized Priors for Facial Appearance Editing
  - High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies
- 考察
  - 3DMM
  - データセット
- 参考文献

論文紹介

RGBD2:　Generative Scene Synthesis via Incremental View Inpainting using RGBD Diffusion Models

目的

- 疎なRGB-Dデータから、シーンの形状と色を復元した動画を生成する
- 入力RGB-D画像から逆投影し、中間メッシュを構築することで、別の視点からのシーンも生成できる

特徴

- 入力データに含まれないシーンを生成は、中間メッシュのレンダリング結果を拡散モデルを用いたインペインティングネットワークによって行う

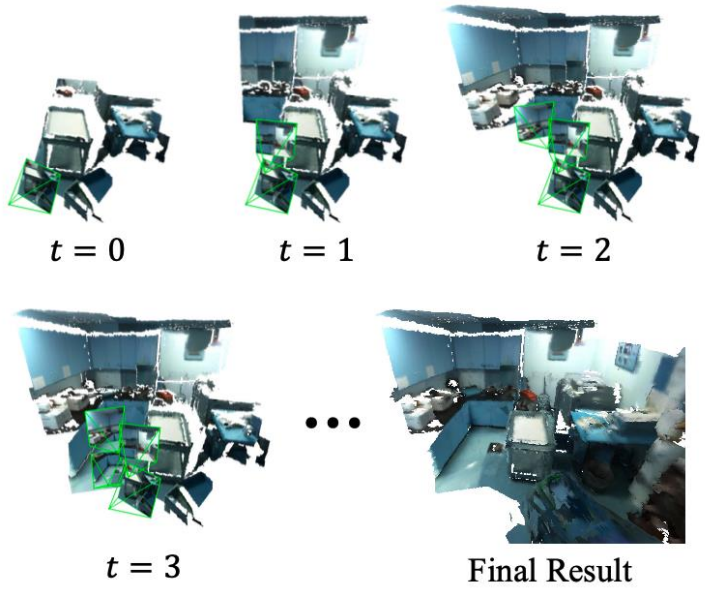

図. モデルの生成過程[1]

論文紹介

RGBD2: Generative Scene Synthesis via Incremental View Inpainting using RGBD Diffusion Models

- **全体アーキテクチャ**
  - 可視領域をマスクにより保持しながら、欠落している視覚的外観およびジオメトリ詳細を合成
  - 逆投影・メッシュ融合・メッシュレンダリングを組み合わせることで、時間フレーム間でのグローバルな3D一貫性を実現

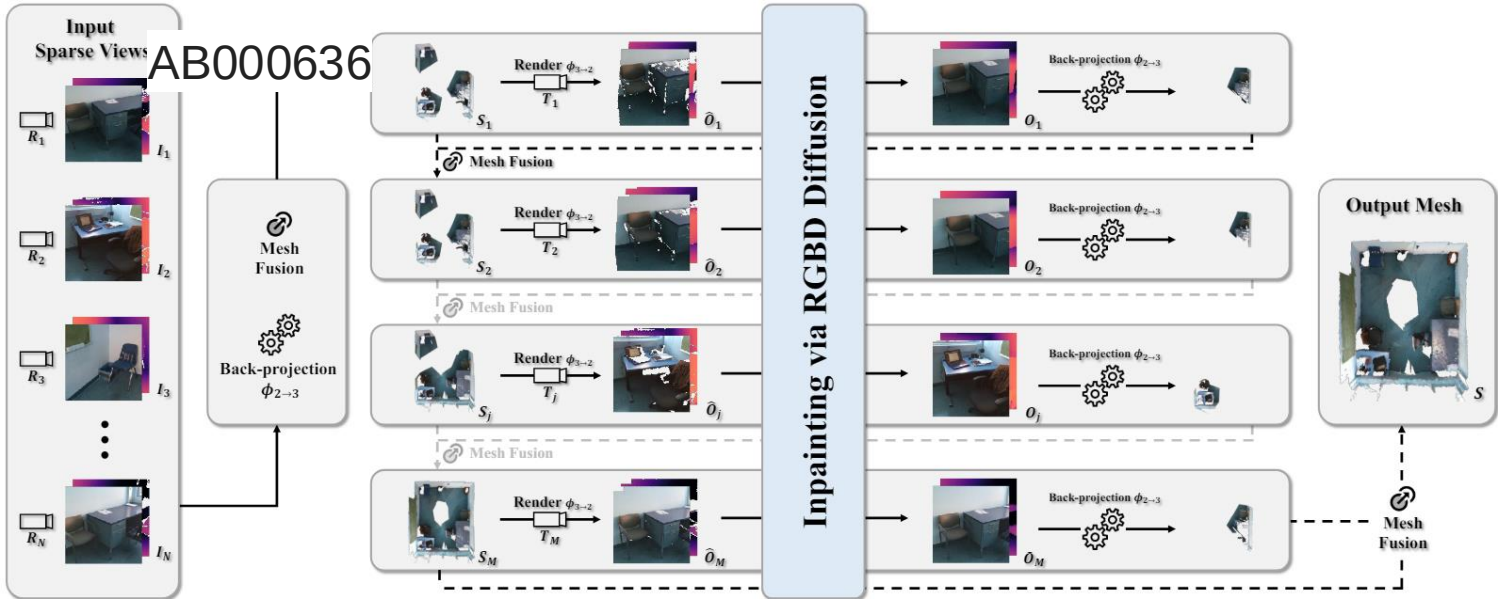

図. RGBD2の生成ネットワーク[1]

論文紹介

RGBD2: Generative Scene Synthesis via Incremental View Inpainting using RGBD Diffusion Models

- **拡散モデル**
  - 可視性に基づいてレンダリングされたバイナリマスクを用いて、欠落部分をインペインティングしたい
  - 制御性を高めるために、無条件モデル$\theta(x_t, c, t)$と条件付きモデル$\theta(x_t, \hat{x}_0, t)$を組み込む。

$$\tilde{\epsilon}_\theta(\mathbf{x}_t, \hat{\mathbf{x}}_0, t) = \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t) + \beta \times [\epsilon_\theta(\mathbf{x}_t, \hat{\mathbf{x}}_0, t) - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)]$$

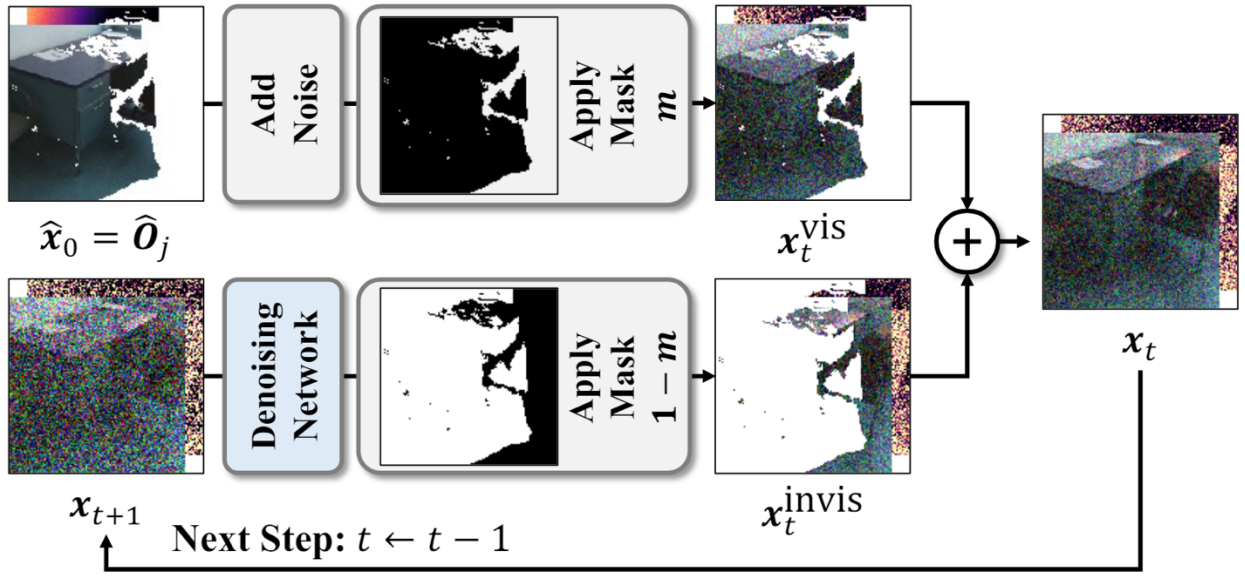

図. RGBD2の生成ネットワーク[1]

論文紹介

RGBD2: Generative Scene Synthesis via Incremental View Inpainting using RGBD Diffusion Models

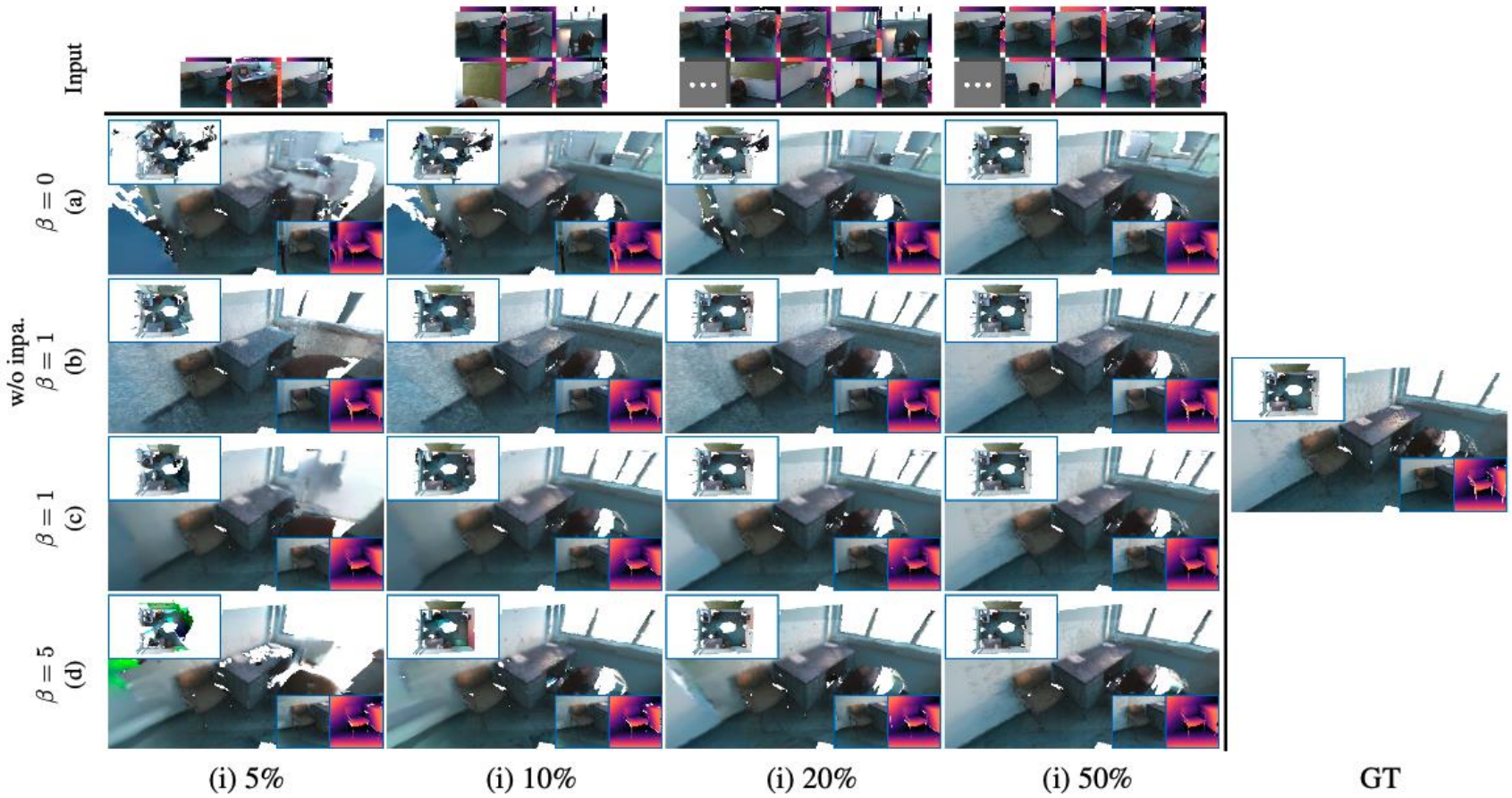

図. 条件付け係数と入力データを変化させた場合の出力[1]

論文紹介

RGBD2: Generative Scene Synthesis via Incremental View Inpainting using RGBD Diffusion Models

- **出力結果の比較**
  - 入力データが増加するとVisual・Geometricの両方の精度が上昇
  - Geometric観点では、入力ビューの割合が大きくなると、最適なβ値も増加
  - 無条件モデル(β=0)を用いると、入力データに含まれるデータの整合性が低下し、β値が大きすぎると視覚的な非現実性が大きくなってしまう(色の過飽和)

Guidance Factor β	Visual												Geometric											
	$PSNR_{color}$				$SSIM_{color}$				$LPIPS_{color}$				MSE_{depth}				CD_{mesh}				$Comp._{mesh}@0.1m$			
	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%
0.0	12.4	14.7	16.5	17.9	0.411	0.496	0.557	0.583	0.520	0.446	0.393	0.359	1.001	0.837	0.761	0.730	1934	850	443	149	0.600	0.781	0.881	0.931
0.5	13.2	15.5	17.1	18.2	0.452	0.530	0.578	0.596	0.496	0.418	0.374	0.347	0.999	0.845	0.772	0.719	1980	606	223	111	0.653	0.818	0.894	0.933
1.0	**14.6**	**16.0**	17.4	18.4	0.522	0.555	0.593	0.603	0.448	0.399	0.359	0.338	**0.825**	0.805	0.688	0.628	**1058**	902	156	100	0.747	0.839	0.909	0.936
2.0	14.5	15.8	**17.5**	**18.4**	**0.532**	**0.561**	**0.598**	**0.606**	**0.439**	**0.393**	**0.352**	**0.334**	0.894	**0.800**	**0.654**	0.593	1562	**515**	**144**	87.2	**0.753**	**0.846**	**0.910**	**0.936**
5.0	13.3	14.9	17.1	18.2	0.488	0.531	0.579	0.598	0.475	0.418	0.367	0.342	0.992	0.856	0.663	**0.582**	2551	1676	175	**87.2**	0.747	0.842	0.908	0.934

図. 条件付け係数と入力データを変化させた場合の出力[1]

論文紹介

RGBD2: Generative Scene Synthesis via Incremental View Inpainting using RGBD Diffusion Models

図. 他モデルとの比較[1]

考察

- 3Dを対象とした再構成を行うタスクでは、RGB-Dデータをそのまま扱うのではなく、メッシュを用いた手法が効果的

- 拡散モデルを用いて生成データに条件付けをする場合は、クラス分類器なし手法を用いたほうが自由度が高く、クラス分類エンコーダの性能に依存しないが、条件付けの強さによって学習の不安定さが生じる可能性がある

- 頭部完全モデルを目指すのであれば、メッシュ化工程(3DMM)を挟む必要がある

参考文献

1. Jiabao Lei, Jiapeng Tang, Kui Jia, “RGBD2: Generative Scene Synthesis via Incremental View Inpainting using RGBD Diffusion Models”, https://arxiv.org/pdf/2212.05993, CVPR 2023, 2023-03-17
2. Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, “High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies”, https://arxiv.org/pdf/2010.05562, 2021-01-29

IPoD: Implicit Field Learning with Point Diffusion for Generalizable 3D Object Reconstruction from Single RGB-D Images

Yushuang Wu, Luyue Shi, Junhao Cai, Weihao Yuan, Lingteng Qiu, Zilong Dong,
Liefeng Bo, Shuguang Cui, Xiaoguang Han

IPoDとは

ポイント拡散を用いたインプリシットフィールド学習で
単一のRGB-D画像からの汎用的な3Dオブジェクト再構築を行うモデル

ポイント拡散：3D点群上の拡散モデル

インプリシットフィールド学習： 3D空間内の各ポイントに値（例えば、距離や密度など）を割り当てる連続的な関数

背景：

- 3D再構成の現在のSOTAは、インプリシットフィールドを学習するTransformerベースのネットワーク
- 教師データが複数の視点から再構築された点群なのでノイズや不完全さが存在する
- 拡散モデルの成長と大規模な3Dデータセットの開発とともに3D拡散モデルが期待されている

取り組んだ問題：

- 2D拡散を3D点群拡散に拡張

- インプリシットフィールドは全ての可能な位置を無目的にクエリし、それらを同等に扱う
  →オブジェクト表面近くのより価値のある局所領域に注目することで細かい形状の詳細を捉える

- インプリシット値の結果を拡散モデルのノイズ除去に用いることで両者の利点を組み合わせる
  →拡散モデルは全体の粗い形状を復元し、
  インプリシットフィールド学習は局所的な細かい詳細について正確な予測を行う

概要

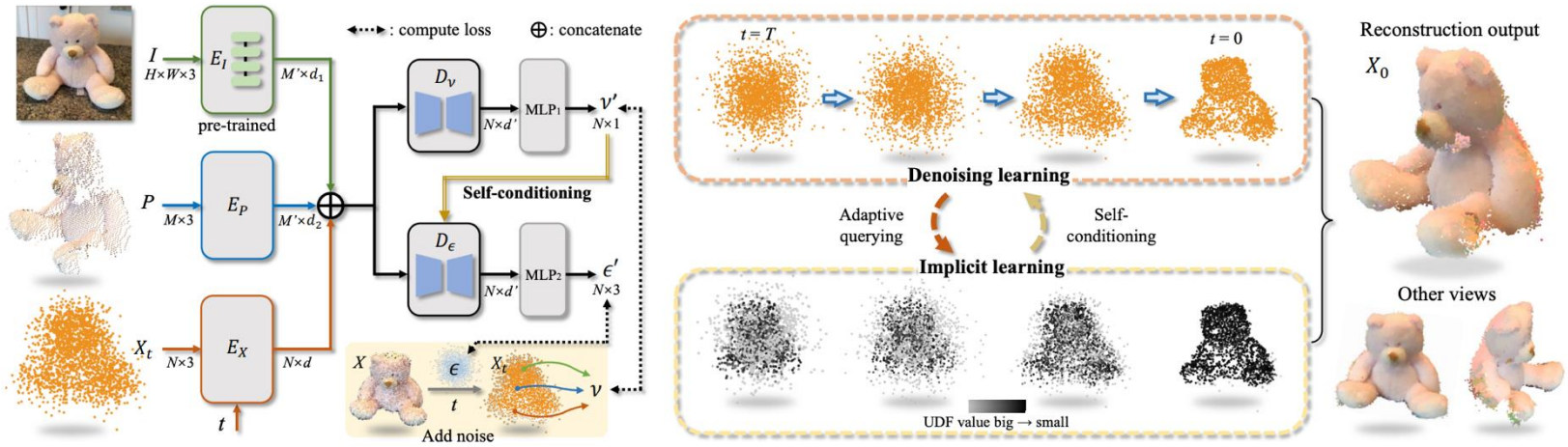

図1. IPoDパイプライン[1]

1. 最初はランダムにサンプリングされたノイズ(X_T)
2. IとPの条件下で、反復的なノイズ除去プロセスt = T, T − 1, ⋯ , 0を経て、徐々に目標形状に近づく
3. ネットワークは適応するX_tに対してインプリシット値を予測

インプリシット値 v:オブジェクト表面までの距離を表すUDF値

$$h_\theta(X_t, t|P, I) \rightarrow (\epsilon, v)$$

h_θの学習

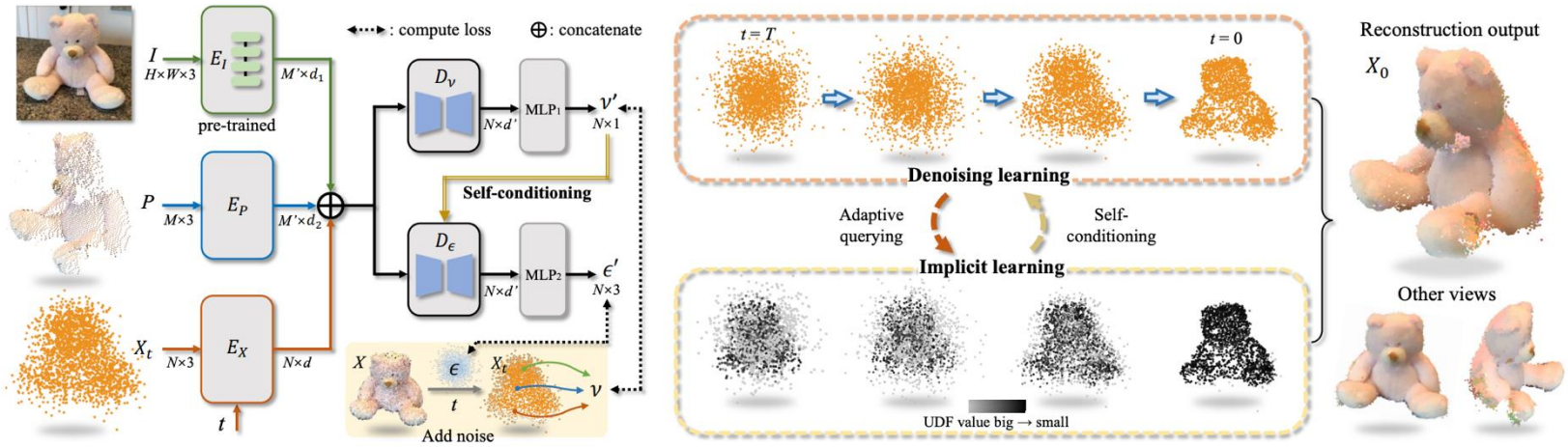

図1. IPoDパイプライン[1]

① $h_\theta(X_t, t|P, I) \rightarrow (\epsilon', v')$

②各反復でノイズϵと時間tをランダムにサンプリング

ノイズの計算：$q(X_t|X_0) = \sqrt{\bar{a}_t}X_0 + \epsilon\sqrt{1-\bar{a}_t}$

vの計算： X_t内の各点がX内の最も近い点までの距離をv

ランダムにサンプリングするよりも効率的に形状を推定できる

③ $L_{uni} = \|v' - v\| + \lambda\|\epsilon' - \epsilon\|$　(λ：重みつけ係数)

インプリシット値 vをRGBに拡張することで色も推定できる

実装

h_θの二つの実装パターン
① Transformerに基づいた実装(上段)
② PVCNNに基づいた実装(下段)

①： Transformerベース
a. 条件画像P, IをViTエンコーダに入力
b. X_tのエンコード
c. 線形層に入力し特徴を得る
d. 類似のアンカー予測でIとPの特徴を
位置と特徴を持つM''のアンカーにエンコード

②： PVCNNベース
✓ 画像特徴をPとX_t内のすべての位置に投影し、拡張して連結

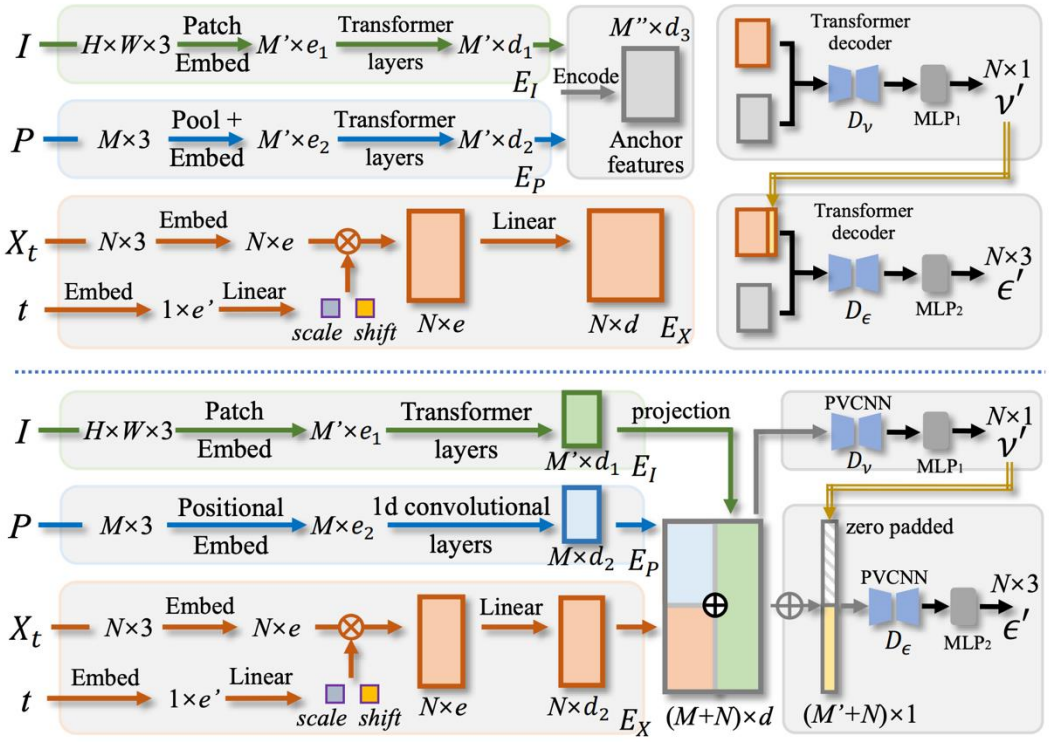

図2. 実装のイメージ[1]

自己条件付け

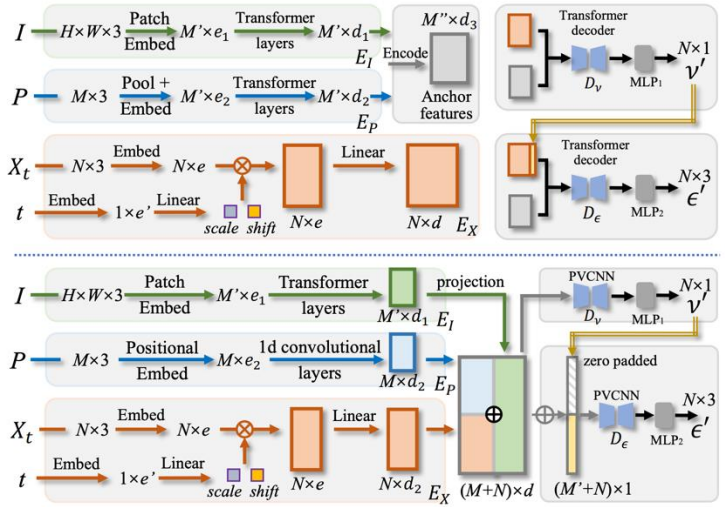

図2. 実装のイメージ[1]

ノイズ除去によるクエリ→細かい形状を生成するためインプリシットフィールドに効果
▶自己条件付けにより可能

今までの手法：
サンプリングプロセス中に直前に生成された変数によって直接条件付け

提案手法：
予測されたインプリシット値v'を自己条件として取ることにより正確な自己条件情報を提案

データセット

CO3D-v2 ：

51のオブジェクトカテゴリの約37k個のビデオ
評価用：10カテゴリ、訓練用：41カテゴリ

MVImgNet：

ゼロショット汎化能力を検査用
238カテゴリにわたる220kのオブジェクトビデオを含む実世界のデータセット
※180° で撮影されているので完全な3D形状のアノテーションはできない

→COLMAPで3Dアノテーション（ 3次元再構成の機能を持ったソフトウェア）

実装の詳細

表1.　学習の詳細

入力画像サイズ	224×224
パッチサイズ	16×16
IとPのエンコード後　特徴量次元	768
拡散モデルステップ数	1000
バッチサイズ　エポック数	64　100
GPU	NVIDIA V100
学習時間	約48時間
学習率	10^{-4}
オプティマイザ	Adam

定量評価

表2. CO3D-v2の結果[1]

Method	Backbone	Acc↓	Comp↓	CD↓	Prec↑	Recall↑	F1↑
PC^2 [35]	PVCNN	0.342	0.214	0.556	24.2	56.2	33.0
PC^2-depth	PVCNN	0.209	0.103	0.312	61.7	87.6	70.7
MCC [61]	Transformer	0.172	0.144	0.316	68.9	72.7	69.8
NU-MCC [28]	Transformer	0.121	0.146	0.266	79.2	84.0	80.9
Ours1	PVCNN	0.163	0.089	0.252	69.0	89.7	76.2
Ours2	Transformer	**0.104**	**0.087**	**0.190**	**85.1**	**90.1**	**87.2**

評価指標：

CD：予測された点群とGTとの間の平均距離
Acc：CDの構成要素
Comp ：CDの構成要素
Prec：予測された点のうち、任意のGT点に対して一定の距離（ρ）以内にある点の割合
Recall：GT点のうち、任意の予測された点に対して一定の距離（ρ）以内にある点の割合
F1： PrecとRecallの調和平均

PC2-depthをCDで19.2%、F値で7.8%向上
NU-MCCをCDで28.6%、F値で7.8%向上

定性評価

図3. CO3D-v2の未確認カテゴリに対する各手法による再構築の結果[1]

今後の研究計画

表6.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