NeurIPS2022
Face Image Synthesis Series

佐藤怜
LINE株式会社
2023/02/28
NeurIPS 2022 論文読み会

Concept: NeurIPS2022の顔画像生成に関する論文をまとめて紹介する

Paper List

1. Controllable 3D Face Synthesis with Conditional Generative Occupancy Fields
2. AniFaceGAN: Animatable 3D-Aware Face Image Generation for Video Avatars
3. FNeVR: Neural Volume Rendering for Face Animation
4. Towards Robust Blind Face Restoration with Codebook Lookup Transformer

Controllable 3D Face Synthesis with Conditional Generative Occupancy Fields

条件付き生成占有場による制御可能な三次元の顔の生成　https://openreview.net/forum?id=Qq-ge2k8uml

問題設定	● 顔画像の集合 (1人1枚，補助情報無)が与えられる．**明示的に表情やポーズを制御できる，写実的な顔画像の生成モデルを獲得したい (図)**
既存手法の問題点	● StyleGAN等は明示的な制御ができない ＆ 内部に3Dモデルを持たないので 3D一貫性がない ● 3D表現を考慮する手法は補助情報 (別視点画像，3Dメッシュ)が必要
アイデアと貢献	● pi-GANをベースに，生成ノイズとしてポーズや表情を与えることで明示的な制御を可能に ＆生成品質の向上のためにcGOFと2つの誤差関数を提案

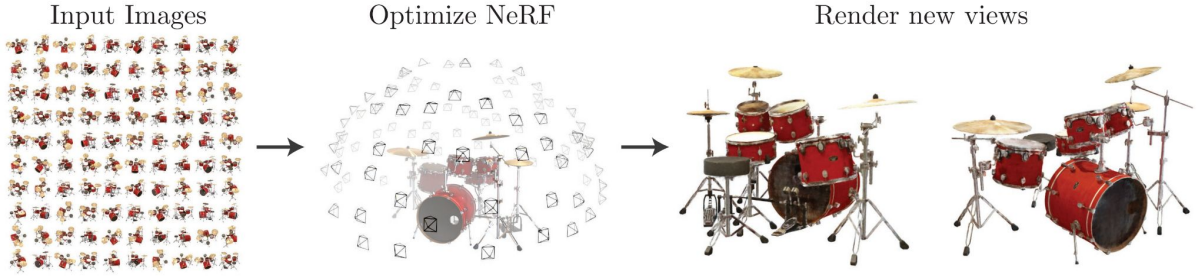
(a) Varying Expressions

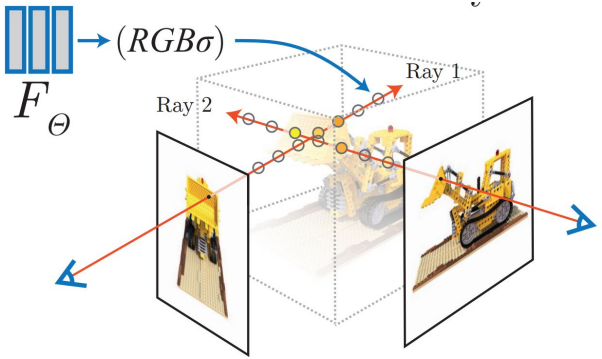
(c) Varying Poses

NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis

視点生成のための Neural Radiance Fieldsとしてのシーン表現 https://arxiv.org/abs/2003.08934 ECCV2020

問題設定	• ある静止したシーンを様々な視点から撮影した画像集合が与えられる．そのシーンの　3Dモデル(形状とテクスチャ)を獲得したい(左図)
前提	• それぞれの画像がどの視点から撮影されたものなのかの情報は　SfM(structure from motion, 複数の画像から3D構造を推定するソフトウェア)を利用して推定する (e.g. COLMAP)
手法	• xyz座標とyaw，pitch角の計5次元の入力から，RGB色とvolume density(不透明度のようなもの)を推定するNNを学習する • 投影面(2Dの画像)上の特定の位置の色は，ray(視線)上でのRGB色をvolume densityで重み付けて足し合わせることで計算できる．**レンダリングした画像から NNのパラメータまでの勾配が計算可能 =微分可能レンダリング(右図)** • データセットの画像の視点でレンダリングを行い，差分を最小化することで　NNを学習する



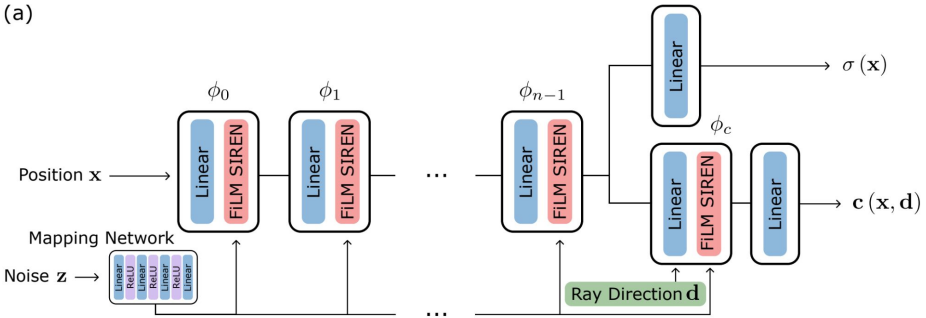


pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis

3Dを考慮した画像生成のための pi-GAN https://arxiv.org/abs/2012.00926 CVPR2021

問題設定	● 画像集合が与えられる．生成品質と 3D一貫性を保った生成モデルを獲得したい (上図)
既存手法の問題	● StyleGAN等は3D一貫性に欠ける ● 3D情報を取り入れた既存手法は，生成品質が良くない
アイデアと貢献	● **NeRFのNNを，xyz座標，視点方向の 5次元に加えて，ノイズを受け取るよう拡張する (下図)** ● 微分可能レンダリングで生成した画像を識別器に入力し敵対的学習

(a)



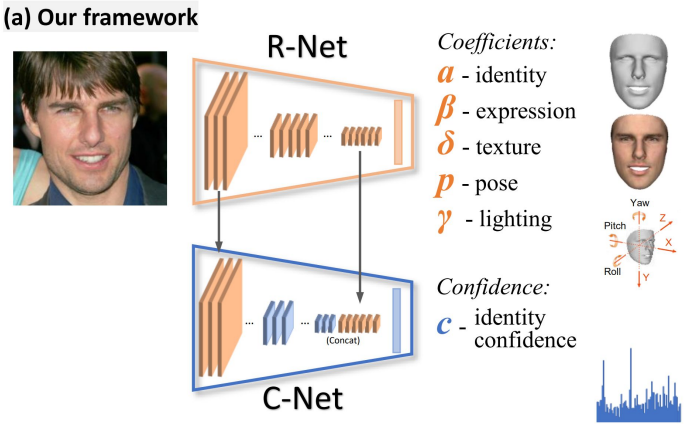
Controllable 3D Face Synthesis with Conditional Generative Occupancy Fields

条件付き生成占有場による制御可能な三次元の顔の生成　https://openreview.net/forum?id=Qq-ge2k8uml

工夫1: 明示的な顔の制御を可能にする

既存手法の利用	• 3DMM: パラメータzを入力すると3Dメッシュを生成する • 1枚の2D画像から3DMMのパラメータを推定する pre-trainedモデルRを利用する(右図) https://arxiv.org/abs/1903.08527
学習	1. 学習データセットから抽出された 3DMMパラメータの分布を多変量正規分布で近似する 2. 近似した分布から生成したパラメータ zと生成した視点 ξから，pi-GANを用いて顔画像を生成する : G(z, ξ) 3. これをRに入力し，3DMMパラメータを求める (hat{z}). これとzとの距離を最小化する (左式)
うれしさ	• **3DMMのパラメータz(解釈可能)によって生成画像を条件付けることができる**

$$\mathcal{L}_{\text{recon}} = \|\hat{\mathbf{z}} - \mathbf{z}\|_1, \quad \text{where} \quad \hat{\mathbf{z}} = \tau(R(G(\mathbf{z}, \xi))).$$

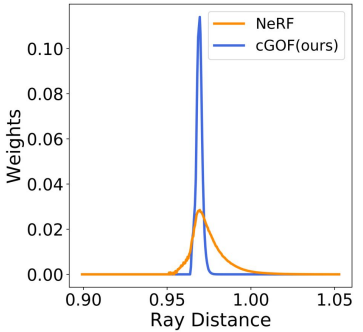


Controllable 3D Face Synthesis with Conditional Generative Occupancy Fields

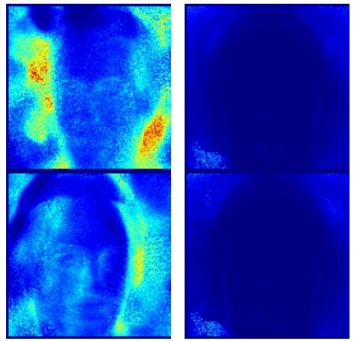
条件付き生成占有場による制御可能な三次元の顔の生成　https://openreview.net/forum?id=Qq-ge2k8uml

工夫2: 生成品質の向上(cGOFの導入)

NeRFの課題	● NeRFによる形状推定は表面の形状があいまいになりがち
仮説と提案	● そもそも顔は不透明なので，表面以外の volume densityは0に近いはず ● 上記の仮説を制約として取り入れた conditional Generative Occupancy Fields(cGOF)を提案する(図)
手順	1. 3DMMのパラメータを先述の分布から生成して， 3Dメッシュを取得する 2. レンダリングする際，顔表面 (1で計算した3Dメッシュ)とrayとの交点に近い部分だけをサンプリングしてレンダリングする **a. 顔の表面付近だけ学習が進んで volume densityが大きくなる** **3. 表面から遠い点については volume densityが0に近くなるように正則化を入れる**



weight w_i distribution



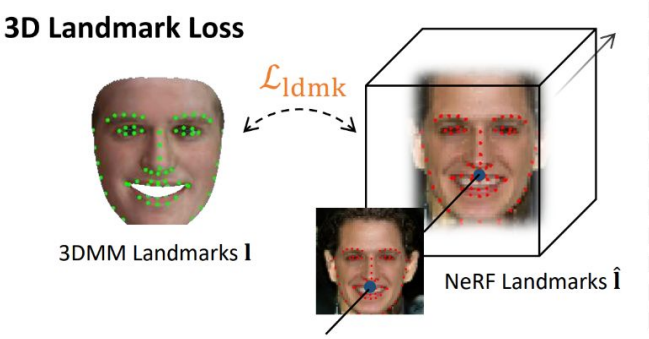
σ_{w_i}(NeRF)　σ_{w_i}(our cGOF)

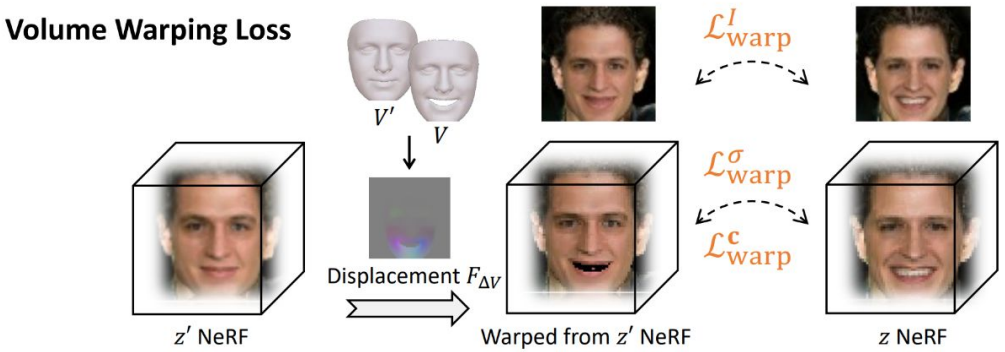
Controllable 3D Face Synthesis with Conditional Generative Occupancy Fields

条件付き生成占有場による制御可能な三次元の顔の生成　https://openreview.net/forum?id=Qq-ge2k8uml

工夫3: 生成品質の向上 (2つの学習誤差の導入)

3D Landmark Loss	1. l: あるzから生成した3Dメッシュのランドマーク座標 2. hat{l}: zにより顔をレンダリングし，ランドマーク推定モデルで推定した 2D上のランドマーク座標を 3Dメッシュ(顔表面)に射影した座標 3. **ランドマークの知覚的な距離で正則化する** ために，lとhat{l}の距離を最小化する (左図)
Volume Warping Loss	1. ある人物z_shapeのある表情z_expについて3DMMでメッシュを生成する 2. 同じ人物の別の表情 z'_expでも同じくメッシュを生成する 3. displacement map F_ΔV: 2つのメッシュについて，投影面の法線ベクトル方向での形状差分を計算する 4. (z_shape, z_exp)で生成したdensityと色をF_ΔVで移動させる 5. **(z_shape, z'_exp)で生成したdensityと色は，4で生成したものに近いはず** -> 近付ける誤差 (右図)

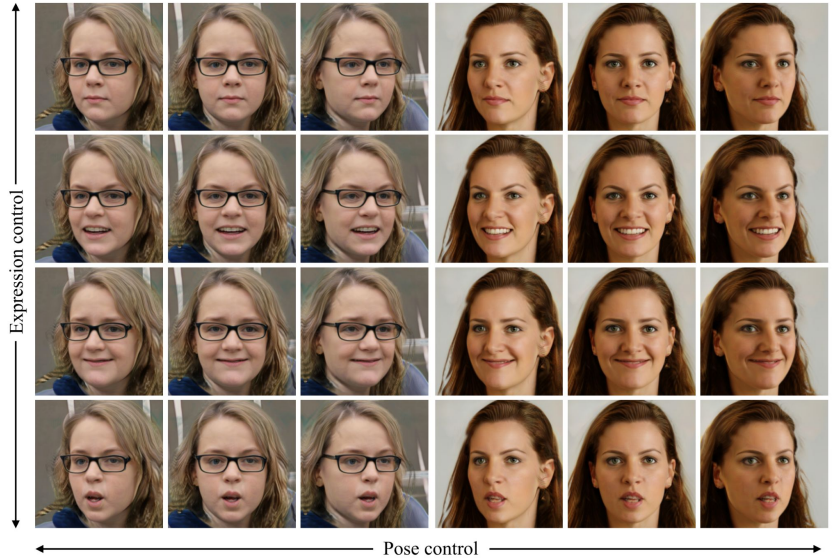




AniFaceGAN: Animatable 3D-Aware Face Image Generation for Video Avatars

ビデオアバターのための 3Dを考慮したアニメータブルな顔画像の生成　https://openreview.net/forum?id=LfHwpvDPGpx

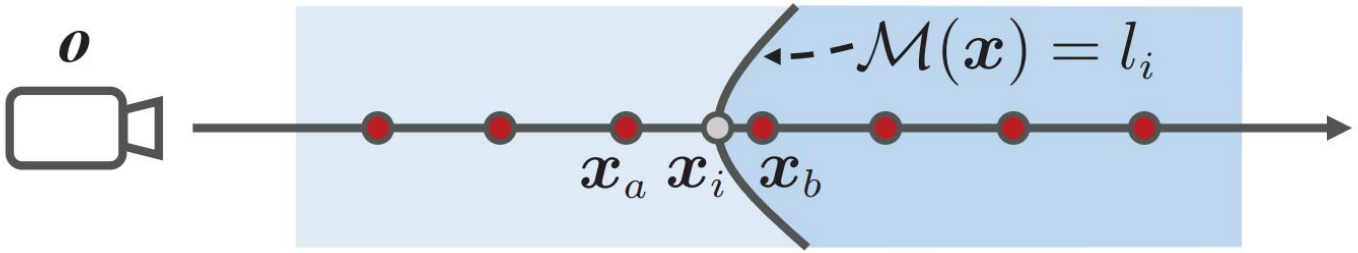
問題設定	● 顔画像の集合 (1人1枚，補助情報無)が与えられる．明示的に表情やポーズを制御できる，写実的な顔画像の生成モデルを獲得したい (図)
既存手法の問題点	● StyleGAN等は明示的な制御ができない ＆ 内部に3Dモデルを持たないので 3D一貫性がない ● 3D表現を考慮する手法は，明示的な表情やポーズの制御が困難
アイデアと貢献	● **NeRF-NNを2つに分割し，それぞれの役割に適した誤差関数を定義して学習する**



GRAM: Generative Radiance Manifolds for 3D-Aware Image Generation

3次元を考慮した画像生成のための GRAM https://arxiv.org/abs/2112.08867 CVPR2022

問題設定	• 3D一貫性のある画像生成モデルを獲得したい
既存手法の問題点	• NeRFはボリュームレンダリングに際してサンプリングする点の数が多いため，学習やレンダリングに時間がかかる
アイデアと貢献	• スカラー場(座標からスカラを出力する)としてのNNを学習する • この出力を用いてisosurfaceを計算し，**物体表面だけをサンプリングして学習効率を上げる (図)** • スカラー場NNの学習はNeRFの学習に用いる勾配を chain ruleで伝播するため，タスクと一貫して学習できる

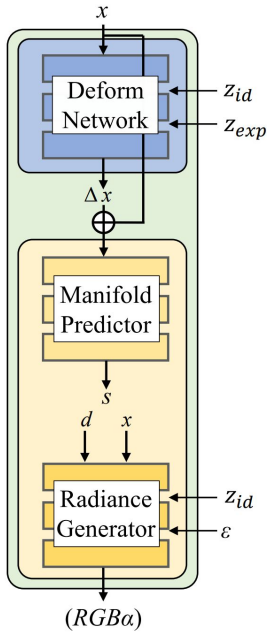


AniFaceGAN: Animatable 3D-Aware Face Image Generation for Video Avatars

ビデオアバターのための 3Dを考慮したアニメータブルな顔画像の生成 https://openreview.net/forum?id=LfHwpvDPGpx

工夫1: NeRF-NNの二分割

Template radiance field	● 3D座標x, z_id(人物のidentity code), ノイズε, 視線dを入力し, RGB色とvolume densityを出力する: G(x, z_id, ε, d) ● 学習メカニズムは GRAMを踏襲 ● きもち: ある人物z_idに共通の3Dモデル(真顔)
Expression-driven 3D deformation field	● 3D座標x, z_id, z_exp(表情パラメータ)を入力し, template radiance fieldとの座標の差分ベクトルを出力する : F(x, z_id, z_exp) ● きもち: ある人物をある表情に補正する 3Dモデル
Image rendering	● templateの出力にdeformationの差分を足し合わせてレンダリングし, 学習する ● **NeRF-NNを二つの役割の異なるモデル (真顔+表情)に分割することが肝(図)**



AniFaceGAN: Animatable 3D-Aware Face Image Generation for Video Avatars

ビデオアバターのための 3Dを考慮したアニメータブルな顔画像の生成　https://openreview.net/forum?id=LfHwpvDPGpx

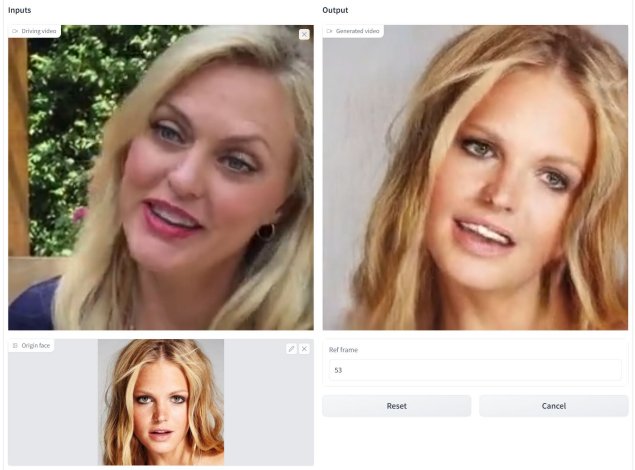
工夫2: 様々な誤差関数の導入

Dense geometry imitation	対象: 両モデル	● 3DMMに(z_id, z_exp)を入力して得たメッシュと，モデルの出力するメッシュの差を最小化
3D landmark imitation	対象: 両モデル	● 以下の2つの3Dランドマークどうしの距離を最小化 ([注]3DMM: (z_id, z_exp)->3Dメッシュは微分可能) a. 3DMMに(z_id, z_exp)を入力して得たメッシュ上のランドマーク b. (z_id, z_exp)で生成した画像から推定した (hat{z_id}, hat{z_exp})を用いて3DMMで生成したメッシュ状のランドマーク
Deformation imitation	**対象: 表情モデル**	● 3DMM上で異なるz_expを用いて生成したときの形状差分と， deformation fieldの出力する形状差分を近付ける
Deformation regularizations	**対象: 表情モデル**	● deformation fieldの出力が小さく，滑らかになるように正則化する

FNeVR: Neural Volume Rendering for Face Animation

顔アニメーションのためのニューラルボリュームレンダリング　https://openreview.net/forum?id=7HTEHRMlxYH

問題設定	● **source image Sとdriving video frames (D_1, D_2, …, D_N)が与えられる．人物的特徴を Sから，モーション(ポーズ)や顔の表情をDから抽出してtalking head video(Nフレームの画像群)を生成するモデルを獲得したい** ● 学習時には複数の driving video framesの系列が与えられる
既存手法の問題点	● 写実性，人物の一貫性といった生成品質に欠ける
アイデアと貢献	● FNeVRを提案して生成品質を改善



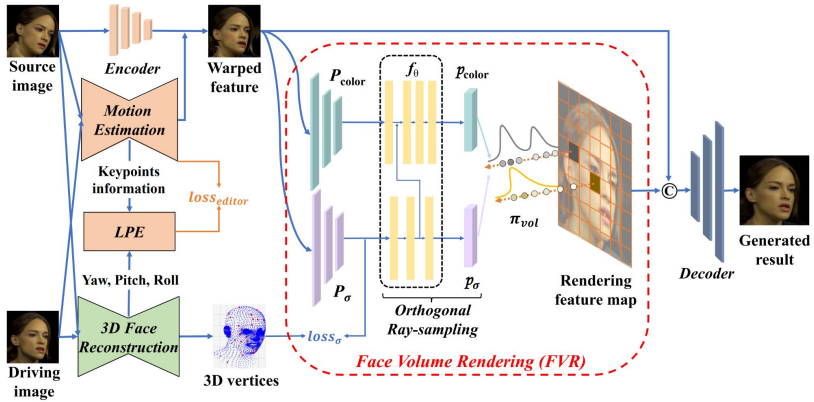
Hugging Face Demo(左図)
https://huggingface.co/spaces/PascalLiu/FNeVR_demo

FNeVR: Neural Volume Rendering for Face Animation

顔アニメーションのためのニューラルボリュームレンダリング　https://openreview.net/forum?id=7HTEHRMlxYH

工夫: 4つのステージからなる生成パイプライン (図)

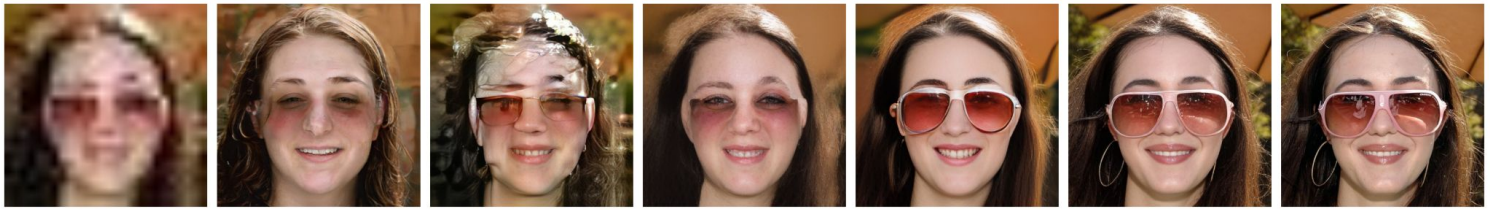
Warped Feature	● SとDのkey pointを検出してmotion field(key pointの移動ベクトルに関するベクトル場)を計算する ● motion fieldを用いてSのfeatureを平面的に移動させて，Dのmotionへと変化させる: warped feature
3D Mesh	● 2D画像からFLAME(3DMM)のパラメータを推定する．FLAMEが推定したパラメータから 3Dメッシュを生成する
3D Feature	● より写実的な生成を行うために，warped featureに奥行の情報を付与する NNを3Dメッシュを用いて学習する
Ray-Sampling	● このタスクはrayが固定なので，xyzだけを入力に受け取って volume densityとcolorを返せば十分 ● そこで，3D featureを受け取り，WxHxDサイズの volume densityとcolorを出力するNNを適用



Towards Robust Blind Face Restoration with Codebook Lookup Transformer

Codebook Lookup Transformerによる頑健な顔復元 https://openreview.net/forum?id=XdDl3bFUNn5

問題設定	• 高品質な顔画像集合が与えられる．**低品質(低画質，欠損有)な画像を高品質(高画質，欠損無)な画像に変換するモデルを獲得したい**
既存手法の問題点	• 低品質から高品質への変換は自由度 (不確実性)がある不良設定問題で，これをうまく制約できていないため，変換結果が忠実でない (図)
アイデアと貢献	• VQVAEで離散的なコードブックを獲得し，この組み合わせで高品質画像を生成することで，うまく自由度を制限



(c) Input (d) PULSE [25] (continuous) (e) GFP-GAN [36] (continuous) (f) Nearest Neighbor (discrete) (g) CodeFormer (discrete) (h) Code GT (discrete) (i) Ground Truth

Towards Robust Blind Face Restoration with Codebook Lookup Transformer

Codebook Lookup Transformerによる頑健な顔復元 https://openreview.net/forum?id=XdDl3bFUNn5

Stage1(上図)	1. 高品質なデータセットで VQVAEを学習する
Stage2(下図)	1. 高品質なデータセットの各画像に対応するコードブックの組み合わせを記録する 2. 高品質なデータセットから低品質なデータセットを作成する 3. 低品質な画像をエンコーダに入力し，エンコーダの出力を Transformerに入力して，正解となる記録しておいたコードブックの組み合わせを予測させる a. エンコーダの出力が正解に近付くような正則化も導入する b. Stage2ではデコーダを固定し，エンコーダと Transformerだけを学習する