一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SIP2022-28,BioX2022-28,IE2022-28,MI2022-28(2022-05)

識別モデルの判断に関わる特徴の深層画像生成による可視化

白　優志† 中尾　恵† 松田　哲也†

† 京都大学大学院 情報学研究科　〒606-8501　京都市左京区吉田本町
E-mail: †y-haku@sys.i.kyoto-u.ac.jp

あらまし　機械学習モデルの判断根拠を知ることは困難であり，モデルの予測に対して解釈性の高い説明を与えることが求められる．本研究では，識別モデルの判断に関わる特徴を深層画像生成に基づいて可視化する枠組みを構築した．入力画像から類似画像と敵対画像を生成し，それらの差分を可視化する．画像生成に条件付き Variational AutoEncoder を用い，特徴空間におけるデータ分布を考慮した画像変換を目指した．手書き文字認識に用いられる MNIST データセットで提案手法の有効性を確認し，下顎骨再建計画データを用いて腓骨片数の分類における重要な特徴を可視化した．

キーワード　解釈可能性, 識別モデル, 深層画像生成，下顎骨再建

Visualization of Important Features for Classifier Decisions using Deep Image Synthesis

Yushi HAKU†, Megumi NAKAO†, and Tetsuya MATSUDA†

† Graduate School of Informatics, Kyoto University　Yoshida Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
E-mail: †y-haku@sys.i.kyoto-u.ac.jp

Abstract　It is difficult to know the basis for the decisions of machine learning models, and it is necessary to provide a highly interpretable explanation for the model's predictions. In this study, we developed a framework for visualizing features related to the decisions of classification models based on deep image generation. Similar and adversary images are generated from the input images, and the differences between them are visualized. Conditional Variational AutoEncoder was used for image generation, aiming at image transformation considering data distribution in the feature space. The effectiveness of the proposed method was confirmed on the MNIST dataset used for handwritten character recognition, and important features in the classification of the number of fibula fragments were visualized using mandible reconstruction planning data.

Key words　Interpretability, Classifier, Deep image generarion, Mandibular reconstruction

1. はじめに

深層学習は情報学分野で近年めざましい発展を遂げている機械学習手法の一つであり，画像認識や自然言語処理などに応用されている．医用画像における臓器領域の抽出 [1] や診断への応用 [2] も報告されているが，モデルがブラックボックスとなり，ユーザーがモデルの判断根拠を知ることが難しいという問題点が指摘されている．特に医療分野では，高い精度のみならず，モデルが決定に至ったプロセスや判断根拠を明確にすることが望まれる [3]．この課題に対して，機械学習モデルの決定や予測に対して説明を与え，モデルの解釈性の向上を目指す Explainable AI（XAI）技術が注目されている．画像認識における XAI 技術として，モデルが決定の際に注目した特徴を可視化し，視覚的に判断根拠を示す研究 [4] [5] がなされているが，得られた結果にノイズが含まれやすいことや，Convolutional Neural Network（CNN）のような特定のモデルに特化した手法であることが課題となっている．また，入力画像の特徴を保持しつつ，識別モデルによって異なる分類結果を得る類似画像と敵対画像を生成することでその差異を判断根拠として可視化する手法 (Similar Adversarial Learning) [6] も提案されている．しかし，これらはノイズによって識別モデルに判断を誤らせるような画像生成にとどまっており，分類に関わる特徴や解釈性の高い説明を与えるには依然として課題が残っている．

一方，医療現場では，医師の知識や経験を駆使して診断や治療などの医療行為が行われている．Computed Tomography (CT) や Magnetic Resonance Imaging (MRI) によって撮像された患者個人の三次元画像が手術計画の作成や術中支援に幅広く応用されており [7]，医療技術の高度化に伴い，情報システムの利用に

よるさらなる医療プロセスの定量化および効率化が期待されている．例えば，患者自身の腓骨を移植することで下顎骨の再建を行う手術 [8] [9] では，医師が再建に用いる腓骨片数と再建のための配置を患者の下顎骨の形状に基づいて決定する必要があるが，現状では医師個人の主観と経験に頼る部分が大きく，決定プロセスは明らかにされていない．客観的で信頼性の高い手術計画を実現するためには，意思決定に至る根拠を明確にすることが望まれる．この課題に対し，スパースモデリングの考え方を用いて，事前に選出された解剖学的特徴から下顎骨再建における重要な特徴量を抽出する試みがなされている [10] [11]．しかし，この方法では抽出可能な特徴が手動で選出された特徴量候補に限定され，特徴量候補の選出時における客観性の維持も難しい．

本研究では，深層生成モデルによる画像生成に基づいて，識別モデルの判断に関わる特徴を可視化する手法を提案する．二値分類を行うモデルを説明対象として，入力画像と特徴が類似し，かつ，識別モデルによって異なる分類結果が得られるような 2 枚の画像を生成する．生成された 2 枚の画像の差が分類に関わる特徴であると定義し，可視化を行う．提案モデルでは，これら画像の生成に Variational AutoEncoder（VAE）[12] を応用する．VAE は生成モデルに深層学習を取り入れた機械学習モデルの一つである．VAE による画像変換では，入力画像から得られる低次元の潜在変数がガウス分布に従うと仮定し，得られる確率分布から画像を再構成する．潜在変数の変更によって画像内の特徴を連続的に変換しながら新たな画像を生成することが可能となる．本研究では，画像生成に VAE を用いることで，生成画像に豊富な多様性を持たせ，従来手法よりも解釈性の高い画像特徴の可視化を目指す．また，識別モデルの出力値に基づいて生成画像を得ることで，識別モデルの判断が明確な場合と曖昧な場合を区別した可視化を達成する．

提案手法の有効性を検証するために，手書き文字認識の学習に利用される MNIST データセットを用いて二値分類の判断に関わる特徴を可視化する実験を行い，従来手法 [4] [6] との比較を行う．さらに，下顎骨再建に用いる腓骨片数の分類問題を対象に，事前に取得された歯科技工士 1 名による 197 通りの手術計画データに提案手法を適用し，腓骨片数の決定における重要な特徴を可視化を試みる．

2. 提 案 手 法

2.1 提案手法の概要

本研究では，二値分類を行う識別モデルの判断に関わる特徴を可視化した画像特徴を説明マップと名付け，従来手法よりも解釈性の高い説明マップを生成することを目的とする．

Similar Adversarial Learning(SAL) と同様に，元画像を類似画像と敵対画像へ変換することによって説明マップを生成する．まず，SAL による画像変換の考え方を説明する．図 1 (a) に特徴空間における画像変換の概念図を示す．集合 χ_0 および集合 χ_1 をそれぞれ識別モデルによって異なる分類結果となる画像群とする．元画像の属する画像群を元画像群と呼び，類似画像と敵対画像生成時に変換先の目標となる画像群を目的画像群と呼ぶ．集合 χ_0 に属する元画像 x に対して，写像 $G_s : \chi_0 \to \chi_0$ によって生成される画像 $G_s(x)$ を類似画像，写像 $G_a : \chi_0 \to \chi_1$ によって生成される画像 $G_a(x)$ を敵対画像とする．このとき，元画像群 χ_0 に対し，類似画像の目的画像群は χ_0，敵対画像の目的画像群は χ_1 である．画像変換モデル G_s, G_a は元画像 x の特徴を維持しつつ，識別モデルによる分類結果が変わるような画像生成となるように学習を進める．これら画像変換モデルによって生成された類似画像 $G_s(x)$ と敵対画像 $G_a(x)$ は識別モデルによる分類結果に影響を及ぼす特徴のみが異なる 2 枚の画像となることが期待される．類似画像と敵対画像の差 $E(x)$ を説明マップと定義し，式 (1) で表す．

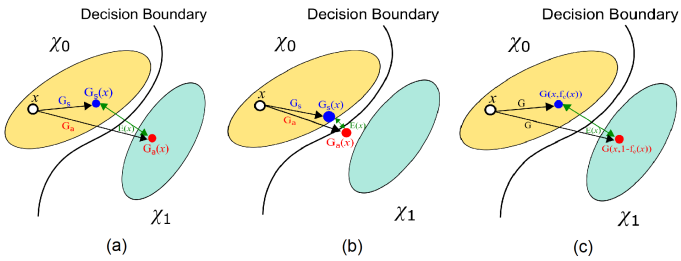

図 1　特徴空間における画像生成の概念．(a) 分布の変換による画像生成，(b)SAL による画像生成，(c) 提案手法による画像生成

$$E(x) = |G_s(x) - G_a(x)| \tag{1}$$

SAL では分類の変換は達成しているものの，図 1 (b) に示すように，χ_0 に属する元画像に対して目的画像群 χ_1 に属するような敵対画像の生成ができないため，解釈性の高い説明マップの生成が難しい．そこで本研究では元画像の特徴を維持しつつ，特徴空間における元画像群と目的画像群間の分布を考慮した画像変換を実現できるような枠組みの提案を目標とする．提案手法の要点は以下の 2 つである．

- VAE による画像変換
- 単一の条件付き画像変換モデル G による画像生成

1 点目について，SAL では画像の生成に U-Net [13] によって構成される AutoEncoder を用いており，分類の変換は交差エントロピーによる損失関数に委ねられている．しかし，この方法で生成される敵対画像はノイズの付加による分類の変換にとどまり，特徴空間において元画像群から目的画像群への変換は達成されない．そこで，入力画像の分布を推定することができる VAE を画像変換モデルに応用することで，特徴空間上での分布に基づく画像変換を目指す．

2 点目について，SAL では異なる 2 つの画像変換モデル G_s, G_a によって類似画像と敵対画像を生成し，L2 ノルムによって各画像の構造の類似度を制約する．しかし，この方法では類似画像と敵対画像の構造が酷似するためダイナミックな変換が難しく，目的画像群への変換ができない．そこで，提案手法では類似画像と敵対画像の生成に VAE による同一の条件付き画像変換モデル G を用いる．条件には識別モデルによる出力を入力画像のラベルとして与え，画像変換モデル G は入力画像とラベルから入力画像の属する分布を推定するように学習する．ラベ

ルの与え方を類似画像と敵対画像が属すべき画像群に応じて変更して与えることで，1 枚の元画像から G が学習した分布上で類似画像と敵対画像を生成する．これにより，元画像の特徴を類似画像と敵対画像の両者に反映しつつ，特徴空間における各データ群の分布間に生成画像が位置するような変換を目指す．

以上の 2 点を踏まえ，図 1 (c) に提案手法が目指す特徴空間上の画像変換の概念を示す．$f_c(x)$ は識別モデル f_c による元画像 x の分類結果の出力とする．提案モデルでは，VAE に条件を与えることができる Conditional VAE(CVAE) [14] を用いて元画像を類似画像と敵対画像に変換し，その差分によって説明マップを生成する．

2.2 Variational AutoEncoder

VAE は変分ベイズ推定法の一つである．AutoEncoder と類似したネットワーク構造を持ち，入力画像の特徴を抽出して潜在変数 z を得るエンコーダと潜在変数 z から画像を再構成するデコーダから構成される．AutoEncoder では潜在変数 z の分布構造が不明であるのに対して，VAE では $z \sim \mathcal{N}(0,1)$ を仮定し，潜在変数 z そのものではなく，潜在変数 z が従う分布の平均 μ と分散 σ を推定するように学習が進められる．VAE の損失関数は KL ダイバージェンスと再構成誤差によって構成される．本研究では，入力画像 $\boldsymbol{x}$，生成画像 $\hat{\boldsymbol{x}}$ に対して再構成誤差を $||\boldsymbol{x}-\hat{\boldsymbol{x}}||_2^2$ とし，CVAE では，VAE に画像のラベル情報を追加的に入力する．

2.3 深層画像変換による説明マップの生成方法

CVAE による説明マップの生成方法について説明する．提案手法は以下の 3 ステップに分けることができる．

STEP1 元画像 x とラベル $f_c(x)$ を用いて CVAE を学習する．
STEP2 学習した CVAE により類似画像と敵対画像を生成する．
STEP3 類似画像と敵対画像の差分によって説明マップを得る．

提案手法における画像生成の流れを図 2 に示す．(a) に STEP1 を，(b) に STEP2 を示す．f_c は説明対象の二値分類用識別モデル，$f_c(x)$ は f_c によって出力される元画像 x の分類結果，G は CVAE による画像変換モデル，$G(x, f_c(x))$ および $G(x, 1-f_c(x))$ は G によって生成される画像を表す．各画像やラベルに示した例は MNIST データセットに対して"7"と"9"の二値分類を対象とした例である．元画像が"7"の場合の例を示しており，類似画像は"7"，敵対画像は"9"となる．まず，STEP1 では類似画像と敵対画像を生成するための CVAE を学習する．入力には元画像 x と識別モデルによる出力 $f_c(x)$ を与え，元画像を再構成するように学習を行う．次に，STEP2 で類似画像と敵対画像を生成する．提案手法では，類似画像を $G(x, f_c(x))$，敵対画像を $G(x, 1-f_c(x))$ と定義する．画像変換モデル G には STEP1 で学習した同一の CVAE を用い，元画像とともに以下のようなラベルを追加的に与える．類似画像は元画像と分類結果を等しくする必要があるため，元画像から識別モデルによって得られる出力値 $f_c(x)$ をラベルとして入力する．一方敵対画像は元画像と異なる分類結果に変換する必要があるため，識別モデルによる出力値 $f_c(x)$ を $1-f_c(x)$ によって反転させた値をラベルとして入力する．ある画像 x に対する識別モデルの出力 $f_c(x)$ が $[y, 1-y]$ の場合，類似画像生成用ラベルは $[y, 1-y]$，敵対画像

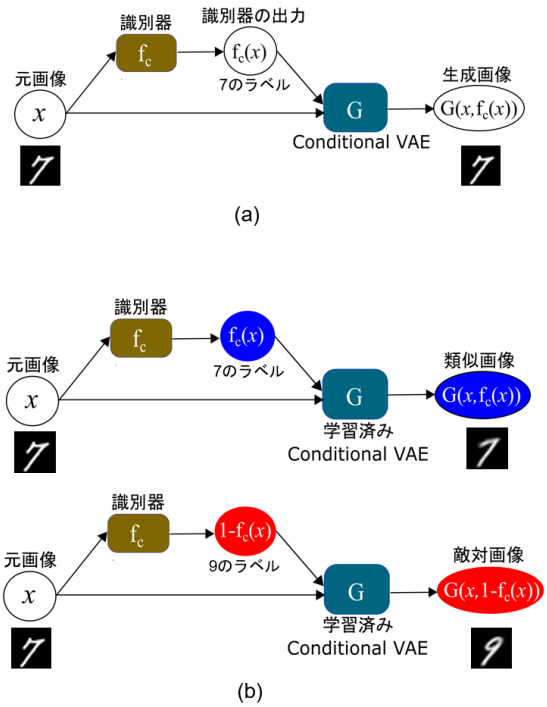

図 2　提案手法の流れ．(a) STEP1，(b) STEP2

生成用ラベルは $[1-y, y]$ となる．STEP3 では，STEP2 で生成される類似画像 $G(x, f_c(x))$ と敵対画像 $G(x, 1-f_c(x))$ から式 (2) によって説明マップ $E(x)$ を算出し，ヒートマップとして入力画像に重ね合わせる．

$$E(x) = |G(x, f_c(x)) - G(x, 1-f_c(x))| \qquad (2)$$

次に，提案手法で用いる CVAE の構造およびラベルの入力方法について説明する．CVAE の構造の一例を図 3 に示す．CVAE では VAE のエンコーダとデコーダの両方に画像のラベルを追加的に入力する．VAE と同様に，エンコーダでは畳み込みとダウンサンプリングによる次元削減を行い，デコーダでは畳み込みとアップサンプリングにより画像を復元する．次に，CVAE へのラベルの入力方法について述べる．本手法では CVAE の構造上，エンコーダとデコーダに対するラベルの入力方法が異なる．エンコーダは画像を入力とするため，ラベルは追加チャネルとして与える．2 次元のベクトルとして得られるラベルで値を統一した入力画像と同じサイズの 2 チャネル画像として入力画像に結合する．一方，デコーダではサンプリングされた潜在変数 z に 2 次元ベクトルのまま直接結合することでラベルを与える．例えば，256×256 ピクセルのある画像 x に対するラベルが $[0.85, 0.15]$ の場合，ラベルの入力は次のようになる．エンコーダには 1 チャネル目の画素値をすべて 0.85，2 チャネル目の画素値をすべて 0.15 とした 256×256 ピクセルの 2 チャネル画像を入力画像に結合することで入力し，デコーダには $[0.85, 0.15]$ を 2 次元ベクトルとして潜在変数に結合することで入力する．

3. 実　　験

本章では MNIST データセットと下顎骨再建計画データを対象に実験を行う．

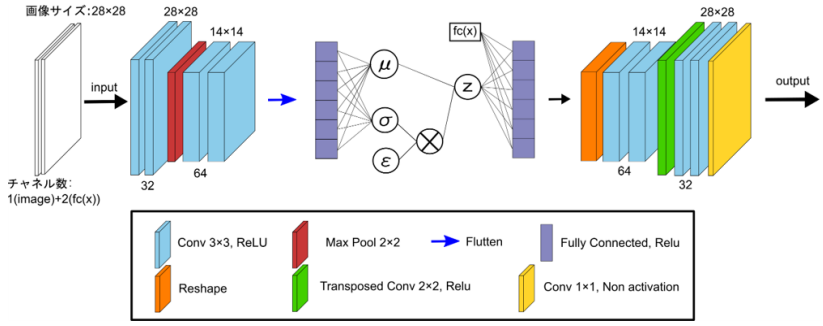

図 3　CVAE の構造例

類似画像と敵対画像の分類精度に基づき，それぞれの実験で個別に CVAE における潜在変数の次元を選定する．本研究では，類似画像と敵対画像が識別モデルによって異なる分類結果となることが必須であるため，識別モデルによる分類精度が高くなる画像生成が可能な次元数を最適な次元数として採用する．

結果については元画像と生成画像の違いおよび説明マップが人間にとって解釈可能なものであるかを目視で評価する．また，元画像に対する識別モデルの出力に閾値を設けて識別モデルによる判断が明確な場合と曖昧な場合を区別し，識別モデルの出力と得られた結果の関係性を検証する．本実験では，2 次元ベクトルで表される識別モデルの出力のうち一方の値が 0.80 を超える例を明確に識別された例とし，0.80 を下回る例を曖昧に識別された例と設定する．例えば，識別モデルの出力が [0.95,0.05] の場合は明確に識別された例となり，[0.33,0.67] の場合は曖昧に識別された例となる．識別モデルには CNN を使用する．なお，本実験は Tensorflow-GPU で実装を行った．

3.1　MNIST データセットによる提案手法の性能評価

本節では，手書き文字認識の学習に利用される MNIST データセットを用いて提案手法の評価実験を行う．画像を入力として，特徴の似た"7"と"9"の二値分類を行う識別モデルを学習し，提案手法を用いて説明マップを生成することを目的とする．提案手法の有効性を検証するために GradCAM，SAL との比較を行う．

3.1.1　実験データと識別モデル

実験で使用するデータセットについて説明する．"7"の画像 6733 枚，"9"の画像 6387 枚の計 13120 枚をトレーニングデータ 12192 枚，テストデータ 928 枚に分割し，2 次元ベクトルを用いて"7"の画像に (1,0)，"9"の画像に (0,1) のラベリングを行った．画素値を [0, 1] の範囲に正規化し，入力画像のサイズは 28×28 ピクセルの 1 チャネルとした．入力画像と定義したラベルを用いて識別モデルを学習した．学習には，学習率 1×10^{-4} の Adam optimizer を使用し，バッチサイズは 32，エポック数は 5 とした．識別モデルによる分類精度は 99.68% となった．

3.1.2　実験条件と結果

3.1.2 節で学習した識別モデルに対する説明マップを提案モデルによって生成した．学習には，学習率 1×10^{-4} の Adam optimizer を使用し，バッチサイズは 32，エポック数は 100 とした．潜在変数の次元数は，類似画像の正解率が 99.89%，敵対画像の正解率が 100% となった 4 次元を採用した．

生成された類似画像と敵対画像の例を図 4 に示す．明確に識別された例を (a) に示し，曖昧に識別された例を (b) に示す．(a) には，1 行目と 2 行目に元画像が"7"の結果を示し，3 行目と 4 行目に元画像が"9"の結果を示す．また，(b) は全て元画像が"9"の結果である．SAL と提案手法はともに元画像の特徴を保持した類似画像生成に成功していた．一方 SAL による敵対画像は線状のノイズが乗った画像にとどまっていた．提案手法による敵対画像は"7"の元画像から"9"の画像生成やその逆の画像生成に成功していたが，類似画像と同様に画像が不鮮明となっていた．

最後に，各手法による説明マップを図 5 に示す．明確に識別された例を (a) に示し，曖昧に識別された例を (b) に示す．(a)，(b) における各行の元画像は図 4 に示す元画像と同様である．まず，明確に識別された例に対して各手法間の比較を行う．GradCAM は広範囲に可視化結果が表れていた．SAL と提案手法について，説明マップを画像中心付近，右上方，左上方の 3 種類の特徴に大別し，それぞれ矢印 A，B，C で示す．元画像が"7"の場合には A や B の部分に特徴が可視化され，"9"の場合には A や B に加えて C の部分にも特徴が可視化された．いずれの場合でも提案手法の方が SAL よりも高い数値を示していた．一方で，提案手法では輪郭の部分に低い数値を示す例が見られた．次に，曖昧に識別された例に着目すると，SAL では明確に識別された例と同様に C に低い数値を示したことに対し，提案手法では明確に識別された例よりも A や B の部分の数値が小さくなり，輪郭付近を中心に特徴が可視化された．

3.2　下顎骨再建計画への適用

本節では，下肢の腓骨を用いて下顎骨の再建を行う下顎骨再建計画データに提案手法を適用する．三次元 CT 画像を下から見た二次元画像をもとに腓骨片数を推定する分類問題を対象に，2 本と 3 本の二値分類における説明マップを生成することを目的とする．

3.2.1　実験データと識別モデル

初めに，手術計画データについて説明する．過去に下顎骨再建術を受けた患者 29 名の頭部および下肢の三次元 CT 画像を対象とし，切除領域による違いを評価するために，Nagai ら [10] [15] によって定義された 6 種類の切除面を使用する．図 6 に示す各切断面は下顎骨の解剖学的特徴に基づいて設定されており，C_0：下顎枝，C_1：正中と C_0 の中点，C_2：C_3 の正中に関する対称点，C_3：正中と C_5 の中点，C_4：C_3 と C_5 の中点，C_5：オトガイ孔となっている．これらの切断面を用いて，(C_0, C_2)，(C_0, C_3)，

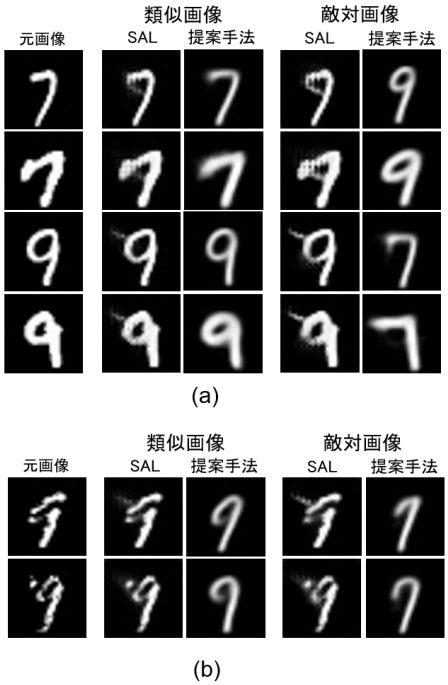

図 4　類似画像と敵対画像の例．(a) 明確に識別された例，(b) 曖昧に識別された例

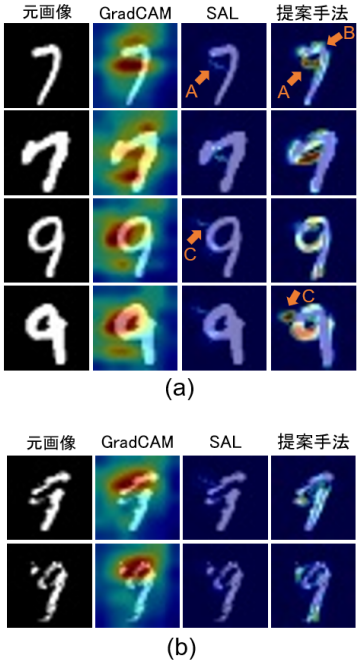

図 5　7 と 9 の分類に対する説明マップ．(a) 明確に識別された例，(b) 曖昧に識別された例

(C_0, C_4)，(C_0, C_5)，(C_1, C_2)，(C_1, C_3)，(C_1, C_4)，(C_1, C_5) の 8 種類の切除領域が各下顎骨データに対して設定されている．患者 29 名の下顎骨データに対してそれぞれ 8 種類の切除領域を設定するため，計 232 例のデータが作成される．このデータに対話型下顎骨再建計画システムを用いて腓骨片数が 1 本，2 本，3 本の場合におけるシミュレーションを行い，歯科技工士 1 名の専門的見地によって最適な腓骨片数とその配置を正解データとして取得している．本節では，腓骨片数 2 本と 3 本の二値分

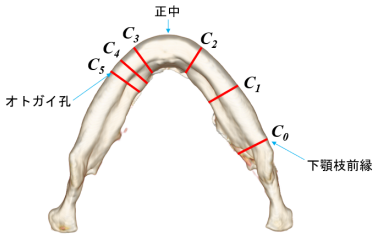

図 6　切除領域を定義する面

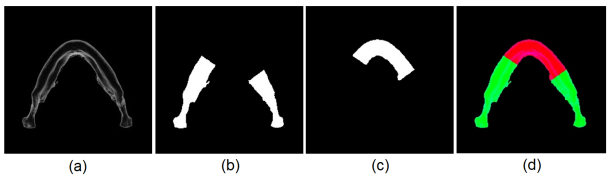

図 7　入力画像．(a) ボリュームレンダリング画像，(b) 残存領域ラベル，(c) 切除領域ラベル，(d) 入力用 3 チャネル画像

類を行う識別モデルを説明の対象として実験を行う．

次に，本実験で使用する入力画像とラベルについて説明する．医師は下顎骨を下から見た二次元画像をもとに腓骨片数と配置を決定することができるという前提の元，三次元 CT 画像を下から見たボリュームレンダリング画像を用いて入力画像を作成した．入力画像は以下の情報を持つ 3 チャネル画像とした．1 チャネル目には医師が実際に見て腓骨片数を決定することができる CT 画像を入力した．図 7 (a) に示すように，ボリュームレンダリング画像をグレースケール化し，背景の画素値を 0 とするためにネガポジ反転を行った。2 チャネル目には，切除後に残る領域の画素値を 255，それ以外の画素値を 0 に持つ二値画像を残存領域ラベルとして入力した．3 チャネル目には，切除領域の画素値を 255，それ以外の画素値を 0 に持つ二値画像を切除領域ラベルとして入力した．case1 の切除領域 (C_1, C_5) における残存領域ラベル，切除領域ラベルを図 7 (b) (c) に示し，3 チャネルの入力画像を図 7 (d) に示す．また，各画像に対する腓骨片数のラベリングは 2 次元ベクトルを用いて，腓骨片数が 2 本のデータを (1,0)，腓骨片数が 3 本のデータを (1,0) とした．

本実験では，全 232 例のうち腓骨片数が 2 本となった 119 例と 3 本となった 78 例の計 197 例を使用する．2 本と 3 本のデータ数のバランスを揃えるため，case1，case3，case9 の 3 患者 20 例をテストデータとし，それ以外の 26 患者 177 例をトレーニングデータに設定した．入力画像と定義したラベルを用い，識別モデルを学習した．学習には，学習率 1×10^{-4} の Adam optimizer を使用し，バッチサイズは 3，エポック数は 50 とした．識別モデルによる分類精度は 90.0% となった．

3.2.2　実験条件と結果

3.2.1 節で学習した識別モデルに対する説明マップを提案モデルによって生成した．学習には，学習率 1×10^{-4} の Adam optimizer を使用し，バッチサイズは 3，エポック数は 30 とした．潜在変数の次元数は，類似画像の正解率が 90.0%，敵対画像の正解率が 95.0% となった 4 次元を採用した

各手法による説明マップを図 8 に示す．上段には 2 本の場合

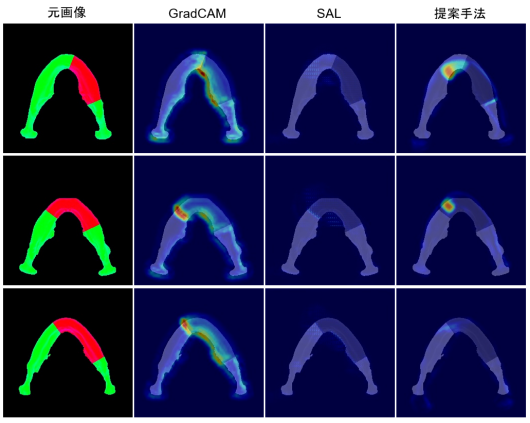

図 8　2 本と 3 本の分類に対する説明マップ．上段：case9，(C_0, C_2)，明確に識別された例 (2 本)，中段：case3，(C_1, C_5)，明確に識別された例 (3 本)，下段：case1，(C_0, C_3)，曖昧に識別された例 (2 本)

における明確に識別された例として case9 の切除領域 (C_0, C_2) を示し，中段には 3 本の場合における明確に識別された例として case3 の切除領域 (C_1, C_5) を示す．また，下段には 2 本の場合における曖昧に識別された例として case1 の切除領域 (C_0, C_3) を示す．GradCAM では，切除領域の端および輪郭に高い数値を示し，下顎骨の輪郭に低い数値を示す説明マップが生成された．SAL による説明マップでは，全ての例で低い数値を示していた．提案手法による説明マップでは，切除領域に特徴が可視化された．特に C_3 から C_5 付近に高い数値が確認され，case9 の切除領域 (C_0, C_2) のように正中付近にも特徴が可視化される例が見られた．また，識別モデルの出力に対する結果の違いを見ると，GradCAM および SAL では説明マップに違いが確認されなかった．一方提案手法による説明マップでは，明確に識別された例については高い数値が確認されたのに対し，曖昧に識別された例については SAL と同様に全体的に低い数値を示しており，識別モデルの出力に応じて説明マップに濃淡の違いが生じることが確認された．

4. おわりに

本研究では，識別モデルの判断に関わる特徴を可視化することを目的に，深層画像変換による説明マップの生成方法を提案した．類似画像と敵対画像の生成に CVAE を採用し，識別対象の画像群の特徴空間内における分布を考慮した変換を達成することで，識別に影響を与える画像特徴の抽出と説明マップの解釈性の向上を目指した．

提案手法の性能を評価するために手書き文字認識の学習に使用される MNIST データセットを用いて"7"と"9"の分類を行う識別モデルに対する説明マップを生成し，従来手法との比較を行った．実験結果から，提案手法によって解釈性の高い説明マップを生成できることが確認された．さらに，提案手法による説明マップでは，識別モデルによる分類の明確さと画像特徴に対する色の濃淡を対応付けられることが確認できた．

また，提案手法を下顎骨再建計画データへ適用し，腓骨片数の分類に重要となる特徴の可視化を試みた．提案手法では計画において医師が重視している切除領域に特徴が可視化された．しかし，提案手法では VAE の性質として生成画像内にぼけが生じる場合があり，特に説明マップ内の輪郭付近の特徴に対する明瞭性を高めることで説明の信頼性や解釈性の向上につながると考えられる．

謝辞

本研究は日本学術振興会 科学研究費補助金 基盤研究 (B) (課題番号：19H04484) の助成による．医用画像の提供，及び，手術計画データ作成に多大なご協力を頂いた洛和会音羽病院口腔外科，今井裕一郎氏，奈良県立医科大学口腔外科，上田順宏氏，畠中利英氏に御礼申し上げます．

文　献

[1] E. Gibson and F. Giganti and Y. Hu and E. Bonmati and S. Bandula and K. S. Gurusamy and B. R. Davidson and S. P. Pereira and M. J. Clarkson and D. C. Barratt, Automatic Multi-Organ Segmentation on Abdominal CT With Dense V-Networks, IEEE Transactions on Medical Imaging, Vol. 37, (2018), 1822-1834.

[2] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature, Vol. 542, (2017), pp. 115–118.

[3] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, What do we need to build explainable ai systems for the medical domain?, ArXiv, Vol.abs/1712.09923, (2017).

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, ICCV , (2017), pp. 618–626.

[5] R. C. Fong and A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, ICCV , (2017), pp. 3449–3457.

[6] M. Charachon, C ´ e. Hudelot, P.H. Courn'ede, C. Ruppli, and R. Ardon, Combining similarity and adversarial learning to generate visual explanation: Application to medical image classification, ICPR, (2021), pp. 7188–7195.

[7] M. Nakao, S. Endo, S. Nakao, M. Yoshida, and T. Matsuda, Augmented endoscopic images overlaying shape changes in bone cutting procedures, PLoS ONE, Vol. 11, No. 9, (2016).

[8] A. F. Flemming, M. D. Brough, N. D. Evans, H. R. Grant, M. Harris, D. R. James, M. Lawlor, and I. M. Laws, Mandibular reconstruction using vascularised fibula., British journal of plastic surgery, Vol. 43, No. 4, (1990), pp. 403–9.

[9] J. S. Brown, C. Barry, M. Ho, and R. Shaw, A new classification for mandibular defects after oncological resection., The Lancet. Oncology, Vol. 17, No. 1, (2016), pp. e23–e30.

[10] 永井一希, 中尾恵, 上田順宏, 今井裕一郎, 畠中利英, 桐田忠昭, 松田哲也, 下顎骨再建に重要な特徴量群抽出に基づく手術計画モデルの生成, 電子情報通信学会技術報告 (MI), Vol. 120, No. 431, (2021), pp. 29–34.

[11] 畑山侑介, 永井一希, 中尾恵, 松田哲也, 下顎骨再建計画に重要な特徴量の複数医師間の解析, 電子情報通信学会技術報告 (MI), Vol. 120, No. 431, (2021), pp. 35–40.

[12] D. P. Kingma and M. Welling, Auto-encoding variational bayes, ICLR, (2014).

[13] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, MICCAI, (2015).

[14] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, Semi-supervised learning with deep generative models, NIPS, (2014).

[15] K. Nagai, M. Nakao, N. Ueda, Y. Imai, T. Kirita, and T. Matsuda, Enumerated sparse extraction of important surgical planning features for mandibular reconstruction, 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), (2020), pp. 5519–5522.