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
SIP2022-27,BioX2022-27,IE2022-27,MI2022-27(2022-05)

画素間の関係性を学習するLambdaNetworkの金属アーチファクト検出への応用

茂森　大亮†　　中尾　　恵†　　松田　哲也†

† 京都大学工学部電気電子工学科　〒606-8501　京都市左京区吉田本町
E-mail: †d_shigemori@sys.i.kyoto-u.ac.jp

あらまし　深層学習に基づく画像変換の金属アーチファクト低減への応用が試みられているが，広範囲に存在するアーチファクトの検出において，離れた画素間の関係性に基づいて特徴抽出を行うことがより効果的であると考えられる．本研究ではLambdaNetworkを適用した領域抽出モデルを用いて金属アーチファクト検出を行った．アーチファクトによる影響を受けた領域を示すラベル画像を含むデータセットを作成し，検出精度の定量評価を試みた．提案手法と従来手法の検出精度を比較し，LambdaNetworkの有効性を検証したので報告する．

キーワード　Self-Attention, U-net，金属アーチファクト検出

Application of LambdaNetwork learning long-range interactions between pixels to metal artifact detection

Daisuke SHIGEMORI†, Megumi NAKAO†, and Tetsuya MATSUDA†

† Undergraduate School of Electrical and Electronic Engineering, Kyoto University　Yoshida Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
E-mail: †d_shigemori@sys.i.kyoto-u.ac.jp

Abstract　Although deep learning-based image transformation has been attempted to be applied to metal artifact reduction, feature extraction based on the relationship between distant pixels is considered to be more effective in detecting artifacts that exist over a wide area. In this study, metal artifact detection was performed using a region extraction model applying a LambdaNetwork. A quantitative evaluation of detection accuracy was attempted by creating a dataset containing labeled images showing areas affected by artifacts. The detection accuracy of the proposed method and conventional methods are compared, and the effectiveness of the LambdaNetwork is verified.

Key words　Self-Attention, U-net, Metal artifact detection,

1. はじめに

機械学習の分野では，様々な画像を対象に物体認識や領域抽出などの研究が行われている．画像を対象とした機械学習において，画像特徴の抽出は重要であり，推定に有効な特徴を効率的に抽出することのできる枠組みが求められる．特徴抽出の代表的な枠組みとして畳み込み演算が知られており，Convolutional Neural Network (CNN) は Graphics Processing Unit (GPU) による並列演算との連携によって画像認識分野に革新をもたらした．画像分類向けに開発されたResNet [1] や，画素単位で分類を行う領域抽出を実現するU-net [2] はいずれも畳み込みによる特徴抽出に基づいた枠組みである．一方，CNNは画像の局所的な特徴を捉えることが可能であるものの，輪郭情報のみを有し，明確な輝度値の勾配や模様を持たない物体に対して鈍感であるという報告がなされている [3]．CNNにおける特徴抽出は，局所領域に対してフィルタ演算を繰り返すプロセスであり，画像内の広範囲に渡る特徴や特徴量間の関係性を十分に捉えられていない可能性がある．この課題に対して，画像の広範囲にわたる特徴を捉える仕組みが研究されている．

近年，自然言語処理の分野で提案されたSelf-Attention [4] の仕組みを画像へと応用した研究 [5] が行われている．Self-Attentionは入力データ間の関係性を考慮する枠組みであり，Self-Attentionを画像へと応用することで，局所領域のみではなく離れた画素間の相互作用を捉えられると考えられている．畳み込み演算を用いずにSelf-Attentionに基づいて画像処理を行うVision Transformer (ViT) [6] が提案されており，画像の分類タスクにおいてCNNを上回る性能を達成することが示されている．しかし，Self-Attentionを用いた学習には，CNNと比較して演算

に要する記憶容量が大きく，高い推定性能を得るためには膨大なデータセットを要する等の問題がある．このような問題を解決するモデルの 1 つとして LambdaNetwork [7] が提案されている．LambdaNetwork は画像全体の特徴と画素間の相対的な位置関係の特徴を抽出する仕組みである．ResNet の畳み込み層を LambdaNetwork に置き換えることで分類精度が向上したという結果が示されており，他分野への応用も期待されている．

一方，臨床医学や生体解析の分野では，医用画像を対象に機械学習の応用が進められている [8]．例えば，CT (Computed Tomography) 画像は診断や手術計画などに役立てられているが，歯科や口腔外科では生体内に歯科金属等が含まれている場合，再構成された画像内にアーチファクトが発生する．このようなアーチファクトは金属アーチファクトと呼ばれ，臨床における診断や手術計画において問題となっている．この問題に対して，深層学習によるアーチファクト領域の検出と画像変換に基づいてアーチファクト低減や画質改善を試みる研究が数多く行われている [9] [10] [12] [11]．金属アーチファクトの画像特徴として，歯科金属を中心として放射状に伸び，比較的画像の広範囲にわたって存在していることが挙げられる．これら画像変換に基づく方法では特徴抽出部に CNN を用いているが，アーチファクトの辺縁部において，十分なアーチファクトの低減に至らない場合がある．CNN による特徴抽出の限界から，広範囲に多様な画像特徴を有する金属アーチファクトを必要十分に抽出することができていない可能性がある．

本研究では，画素間の関係性を学習する新しい機械学習の枠組みを医用画像へ応用し，金属アーチファクトの検出性能を調査する．上述の通り，歯科金属による金属アーチファクトは画像内に広範囲にわたる形状特徴をもつが，画素間の相互関係を学習することによって，検出精度の向上が見込める．画素間の相互関係を学習するための枠組みとして LambdaNetwork を採用し，畳み込み演算による領域抽出アーキテクチャである U-net に対して，エンコーダにおける一部の畳み込み層 (Convolution Layer) を LambdaNetwork を適用した Lambda Layer に置き換えた枠組みを用いる．

LambdaNetwork の有効性を検証するために，特徴抽出部に LamdaLayer を用いる場合と従来の Convolution を用いる場合の金属アーチファクトの検出精度を確認する実験を計画した．29 症例のアーチファクトを含まない実患者の頭頸部 CT 画像を対象に，歯科金属から発生する金属アーチファクトを複数パターンシミュレートすることよってアーチファクト合成画像を生成した．得られたアーチファクト合成画像とアーチファクト領域を示すラベル画像を用いて教師あり学習を行い，アーチファクト領域の検出精度を確認した．

2. 手　　法

2.1 問題設定

1 章で述べた通り，歯科金属由来のアーチファクトは金属を中心として放射状に，かつ広範囲に広がるという特徴を有する．また，LambdaNetwork は従来の畳み込みによる特徴抽出と異なり，広範囲にわたる画素間の関係性を考慮した特徴抽出が可能であると考えられる．本研究では，広範囲の画素領域に形状特徴を持つ対象の検出において，離れた画素間の関係性を考慮することが有効であるという仮説に基づき，CT 画像における金属アーチファクトを対象に LambdaNetwork を用いて領域検出を試みる．U-net における畳み込み層を LambdaNetwork の層へと置き換えることで，階層化された LambdaNetwork を構築することができる．本ネットワークは Ou らによって提案された LambdaUNet [13] を参考に実装する．LambdaUNet はボリュームデータを対象にしており，2D スライスのみからではなく隣接スライスからも特徴抽出が可能であるが，本研究では金属アーチファクト検出における有効性を確認する最初の試みとして，2D スライス 1 枚を対象とした学習と検出を前提とする．

金属アーチファクトの領域検出精度はダイス係数などの定量評価指標によって求められる．この定量評価指標の算出には画像内におけるアーチファクト領域を示す正解のラベル画像が必要となる．しかし，同一患者からアーチファクトを有する画像と，アーチファクトを有しない画像を取得することは困難であり，またアーチファクトは非常に複雑であるため，アーチファクトによる影響を受けた領域を手動で抽出し，正解ラベル画像を作成することも容易ではない．そこで本研究では，実患者から撮像されたアーチファクトを有しない元画像からシミュレーションによってアーチファクト合成画像を生成し，元画像とアーチファクト合成画像の差分によって正解ラベル画像の作成を試みる．本研究では実患者から撮像されたアーチファクトを有しない 3D 画像を元画像とし，シミュレートしたアーチファクトを有する 3D 画像をアーチファクト合成画像とする．金属アーチファクトのシミュレーションには，従来の CNN を用いたアーチファクト低減の研究 [10] で使用されたサイノグラム逆投影のアルゴリズムを用いる．

2.2 画素間の関係性を学習する LambdaNetwork

本節では，LambdaNetwork により特徴抽出を行う Lambda layer のアルゴリズムについて説明する．LambdaNetwork は Self-Attention の考え方をもとにしたアーキテクチャであるため，その仕組みについて述べるにあたって Self-Attention で用いられてる Query($\boldsymbol{Q}$)，Key($\boldsymbol{K}$)，Value($\boldsymbol{V}$) の 3 つの用語を使用する．LambdaNetwork におけるこれらの役割は Self-Attention とは異なる．以下では，LambdaNetwork における $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ の役割を踏まえてアルゴリズムの流れを説明する．説明するにあたり，LambdaNetwork が学習する関係性について元論文における説明をそのまま用いて，以下のように定義しておく．

- content-based の関係性 : 画素間の相対的な位置関係を除く関係性
- position-based の関係性 : 画素間の相対的な関係性

LambdaNetwork ではこれらの関係性を踏まえて特徴抽出を行う．Self-Attention では $\boldsymbol{Q}$ と $\boldsymbol{K}$ の内積を計算して Attentnion map を生成するが，計算コストを要する．そこで LambdaNetwork では $bm\boldsymbol{Q}$ を直接変換して出力とすることで，計算を効率的に行う．このとき，$\boldsymbol{Q}$ を出力の次元へと変換するための線形写像が必要となるため，これを $\boldsymbol{\lambda}$ とする．入力画像として $\boldsymbol{X} \in \mathbb{R}^{h \times w \times d}$ が与えられた場合を考える．h, w, d はそれぞれ入力画像の縦

幅，横幅，チャネル数である．LambdaNetwork では各ピクセルに対して計算処理を行うため $n = h \times w$ として，入力画像 $\boldsymbol{X}$ を $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{n \times d}, \boldsymbol{x}_i \in \mathbb{R}^d$ として扱う．$\boldsymbol{x}_i$ はラスタ順で i 番目のピクセルを表すチャネル方向のベクトルである．LambdaNetwork のアルゴリズムの概要を図 1 に示す．

最初に式 (1) に基づいて入力画像 $\boldsymbol{X}$ から $\boldsymbol{Q}$ を作成する．

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{Q}} \in \mathbb{R}^{n \times k} \tag{1}$$

k は $\boldsymbol{Q}$ のチャネル数であり，$\boldsymbol{W}_{\boldsymbol{Q}} \in \mathbb{R}^{d \times k}$ は学習可能な線形写像である．$\boldsymbol{X}$ と同様に $\boldsymbol{Q}$ もピクセル単位で表現し，$\boldsymbol{Q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n), \boldsymbol{q}_i \in \mathbb{R}^k$ とする．各 $\boldsymbol{q}_i$ は k チャネルの特徴と i 番目のピクセルであるという位置情報によって特徴づけられる．この $\boldsymbol{Q}$ に対して $\boldsymbol{\lambda}$ を乗じることにより，出力を得る．出力画像 $\boldsymbol{Y}$ のチャネル数を v として $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n), \boldsymbol{y}_i \in \mathbb{R}^v$ とする．$\boldsymbol{q}_i$ を $\boldsymbol{y}_i$ へと変換するには線形写像 $\mathbb{R}^k \to \mathbb{R}^v$ が必要であり，$\boldsymbol{\lambda}_i \in \mathbb{R}^{k \times v} (i = 1, 2, \ldots, n)$ を用意する．出力は式 (2) で表される．

$$\boldsymbol{y}_i = \boldsymbol{\lambda}_i^T \boldsymbol{q}_i \tag{2}$$

$\boldsymbol{\lambda}$ には content-based と position-based の関係性の特徴を集約し，式 (2) に基づいて content-based と position-based の関係性を捉える．

次に，$\boldsymbol{Q}$ を出力 $\boldsymbol{Y}$ へと変換する線形写像 λ を算出するアルゴリズムについて説明する．λ は content-based の関係性と position-based の関係性を考慮した特徴をもつ．線形写像 λ を求めるにあたり入力画像 $\boldsymbol{X}$ を用いる．出力への変換に用いる $\boldsymbol{X}$ と区別するため，λ を算出するために用いる入力画像を Context($\boldsymbol{C} \in \mathbb{R}^{m \times d}$) と定義する．LambdaNetwork における Context は入力画像 $\boldsymbol{X}$ であるとは限らないが，本実験では $\boldsymbol{X} = \boldsymbol{C}$ とする．文字の混在による煩雑さを避けるため Context のチャネル以外の変数には $\boldsymbol{X}$ とは異なる文字を使用し，$m = h \times w$ とする．Lambda Layer では，まず $\boldsymbol{C}$ から式 (3) に基づいて $\boldsymbol{K}, \boldsymbol{V}$ を得る．

$$\begin{aligned} \boldsymbol{K} &= \boldsymbol{C}\boldsymbol{W}_{\boldsymbol{K}} \in \mathbb{R}^{m \times k} \\ \boldsymbol{V} &= \boldsymbol{C}\boldsymbol{W}_{\boldsymbol{V}} \in \mathbb{R}^{m \times v} \end{aligned} \tag{3}$$

$\boldsymbol{W}_{\boldsymbol{K}} \in \mathbb{R}^{d \times k}, \boldsymbol{W}_{\boldsymbol{V}} \in \mathbb{R}^{d \times v}$ はそれぞれ学習可能な線形写像である．$\boldsymbol{K}, \boldsymbol{V}$ はそれぞれ $\boldsymbol{K} \in \mathbb{R}^{h \times w \times k}, \boldsymbol{V} \in \mathbb{R}^{h \times w \times v}$ とし，k チャネルと v チャネルの画像と捉えることができるが，計算を行列の内積で行うため図 1 における $\boldsymbol{K}, \boldsymbol{V}$ では行数を総画素数 m，列数をそれぞれのチャネル数として行列で表現している．

以下では $\boldsymbol{K}, \boldsymbol{V}$ から content-besed の関係性に基づいて特徴抽出を行う流れと position-based の関係性に基づいて特徴抽出を行う流れに分けて説明する．それぞれの特徴を [7] と同じように content lambda, position lambda として定義する．

content lambda($\boldsymbol{\lambda}^c$) は入力画像から content-based の関係性をもとに $\boldsymbol{q}_i$ を出力へと変換する線形写像 $\mathbb{R}^k \to \mathbb{R}^v$ である．$\boldsymbol{K}$ に対して softmax 関数 σ を適用しチャネルごとに正規化したものを $\boldsymbol{\sigma}(\boldsymbol{K})$ とする．$\boldsymbol{\sigma}(\boldsymbol{K})$ と $\boldsymbol{V}$ により $\boldsymbol{\lambda}^c$ は式 (4) で定義される．

$$\boldsymbol{\lambda}^c = \boldsymbol{\sigma}(\boldsymbol{K})^T \boldsymbol{V} \tag{4}$$

より詳細な説明のため，$\boldsymbol{K'} \in \mathbb{R}^{k \times h \times w}, \boldsymbol{V'} \in \mathbb{R}^{v \times h \times w}$ とし，$\boldsymbol{K'} = (\boldsymbol{K}'_1, \boldsymbol{K}'_2, \ldots, \boldsymbol{K}'_k), \boldsymbol{K}'_i \in \mathbb{R}^{h \times w}$ と $\boldsymbol{V'} = (\boldsymbol{V}'_1, \boldsymbol{V}'_2, \ldots, \boldsymbol{V}'_v)$, $\boldsymbol{V}'_i \in \mathbb{R}^{h \times w}$ で表現する．$\boldsymbol{K}'_i, \boldsymbol{V}'_i$ はいずれも $\boldsymbol{C}$ から線形変換によって得られる特徴マップである．まず各 $\boldsymbol{K}'_i$ を softmax 関数によって正規化し，k 個の確率マップを生成する．これを $\boldsymbol{\sigma}(\boldsymbol{K})_{\boldsymbol{i}}, (i = 1, 2, \ldots, k)$ とする．次に各 $\boldsymbol{V}'_1 \sim \boldsymbol{V}'_v$ と $\boldsymbol{\sigma}(\boldsymbol{K})_i$ の要素ごとの積の和を計算する．同様の処理を $\boldsymbol{\sigma}(\boldsymbol{K})_{\boldsymbol{1}} \sim \boldsymbol{\sigma}(\boldsymbol{K})_{\boldsymbol{k}}$ について行い，$\boldsymbol{\lambda}^c \in \mathbb{R}^{k \times v}$ を得る．

以上を踏まえて $\boldsymbol{\lambda}^c$ の解釈について説明する．$\boldsymbol{K'}$ は k 種類の特徴に関する特徴マップであり，各特徴マップに softmax 関数を用いた正規化を行うことは，k 種類の特徴それぞれに関して各画素の重要度合を確率で表現することを意味する．$\boldsymbol{V}'_1 \sim \boldsymbol{V}'_v$ と $\boldsymbol{\sigma}(\boldsymbol{K})_i$ の要素ごとの積の和を計算することで，各画素の重要度に応じて $\boldsymbol{V}_i$ の画素が重み付けられ，加算される．したがって $\boldsymbol{\lambda}^c$ は k 種類の特徴に関して画像全体から抽出される特徴であり，画素間の相対的な位置関係ではなく絶対的な位置における重要度合の関係をもとに入力画像の特徴をまとめている．

次に position lambda($\boldsymbol{\lambda}^p$) について説明する．$\boldsymbol{\lambda}^p$ は入力画像から position-based の関係性をもとに $\boldsymbol{q}_i$ を出力へと変換する線形写像 $\mathbb{R}^k \to \mathbb{R}^v$ である．位置 i における $\boldsymbol{q}_i$ に対しては，位置 i の画素とその他の画素との相対的な位置関係を考慮し $\boldsymbol{\lambda}_i^p$ を生成する．したがって画素数 n と同じ数の $\boldsymbol{\lambda}_i^p (i = 1, 2, \ldots, n)$ を用意する．画素間の相対的な位置関係を考慮するため，学習パラメータである position embedding $\boldsymbol{E}_i$ を用いる．$\boldsymbol{E}_i$ と $\boldsymbol{V}$ によって $\boldsymbol{\lambda}_i^p$ は式 (5) で定義される．

$$\boldsymbol{\lambda}_i^p = \boldsymbol{E}_{\boldsymbol{i}}^T \boldsymbol{V} \tag{5}$$

詳細な説明のため，$\boldsymbol{\lambda}^c$ の時と同様に $\boldsymbol{V'} = (\boldsymbol{V}'_1, \boldsymbol{V}'_2, \ldots, \boldsymbol{V}'_v), \boldsymbol{V}'_i \in \mathbb{R}^{h \times w}$ を考える．画像サイズ $h \times w$ に対して embedding filter$\in \mathbb{R}^{(2h-1) \times (2w-1) \times k}$ を用意する．embedding filter は学習パラメータである．$\boldsymbol{E}_i$ は位置 i の画素とその他の画素との相対的な位置関係を表す k 枚のフィルターである．$\boldsymbol{E}_i$ の各フィルターと $\boldsymbol{V'}$ との要素ごとの積の和によって $\boldsymbol{\lambda}_i^p$ が得られる．

以上をまとめると $\boldsymbol{\lambda}_i^p$ は位置 i における画素とその他すべての画素との相対的な位置関係を考慮し，位置関係に応じて重み付きで特徴抽出を行う．embedding のフィルターは k チャネル存在するため，k 種類の特徴に関して相対的な位置関係に基づいた特徴抽出を学習していると考えられる．

一方で 1 つの画素とその他すべての画素との相対的な位置関係を考慮するフィルターを全画素分用意すると，すべての画素間の位置関係を個別に捉えることが可能となるが，position embedding のフィルター $\boldsymbol{E} \in \mathbb{R}^{n \times m \times k}$ によって計算するため，入力画像のサイズに応じて計算コストが高くなり，記憶容量も必要となる．[7] では各画素と相対的な位置関係を考慮する領域を局所範囲に制限する方法も提案されている．本研究においても記憶容量の都合により位置関係を考慮する領域を局所範囲に限定する．局所領域において相対的位置関係を捉えるために畳み込みを用いる．

以上から content lambda と position lambda の役割は以下のように定められる．

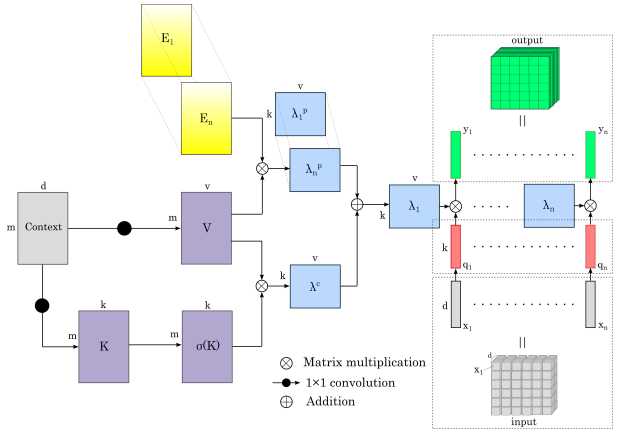

図 1　LambdaNetwork の概要

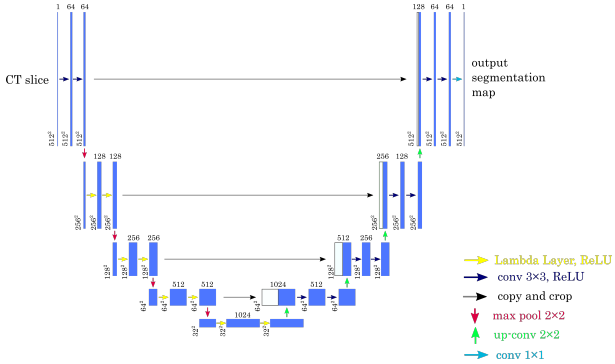

図 2　LambdaNetwork を適用した領域抽出モデル

- content lambda $\boldsymbol{\lambda}^c$：k 種類の特徴に対して各画素の重要性を考慮して抽出する特徴
- position lambda $\boldsymbol{\lambda}_i^p$：k 種類の特徴に対して周辺画素との相対的な位置関係を考慮して抽出する特徴

$\boldsymbol{\lambda}^c$ と $\boldsymbol{\lambda}_i^p$ の特徴をもとに式 (2) に基づいて出力へと変換する．

$$\boldsymbol{y}_i = \boldsymbol{\lambda}_i^T \boldsymbol{q}_i = (\boldsymbol{\lambda}^c + \boldsymbol{\lambda}_i^p)^T \boldsymbol{q}_i \tag{6}$$

$\boldsymbol{\lambda}^c$ は画像の重要な特徴を大局的に捉えるが各画素の位置情報については考慮していない．そこで $\boldsymbol{\lambda}_i^p$ を加えることで相対的な位置関係に関する特徴を考慮する．

2.3　LambdaNetwork を用いた領域抽出モデル

本節では，金属アーチファクトの領域抽出を実現する LambdaNetwork に基づくアーキテクチャについて説明する．本実験で用いる領域抽出モデルは Ou ら [13] によって提案された LambdaUNet を参考にする．モデルの全体像を図 2 に示す．領域抽出モデルとして用いられる U-net のエンコーダの一部において，畳み込み層の代わりに LambdaNetwork によって特徴抽出を行う Lambda Layer を適用する．以降，U-net に LambdaNetwork を適用したモデルを LambdaNetwork モデルと呼ぶ．また，LambdaUNet はボリュームデータを対象にしているが，本実験では 2 次元のスライス画像を対象とする．

本研究では，広範囲にわたる画素間の相互関係を考慮して特徴抽出を行うことがアーチファクトの形状的特徴の把握に有効か否か検証することを目的としている．したがって position lambda を計算する際には，各画素とその他すべての画素との相対的位置関係を考慮することが望ましい．しかし，記憶容量の都合上全画素間について pofsition embedding を行うためのフィルターを用意することは難しい．そこで position lambda を計算する際に相対的位置関係を考慮する範囲を局所領域に制限する．解像度を下げながら局所領域における特徴抽出を繰り返すことで，より広範囲における位置関係を考慮する．position lambda については CNN 同様，離れた画素との相対的位置関係を考慮した直接的な計算は実現できない．一方で content lambda については各 Lambda Layer において入力画像のすべての画素を考慮し，大局的に特徴を捉えることが可能となる．

3.　実　　験

3.1　データセットの作成

まず，本実験で用いるデータセットの作成方法について説明する．本研究では元画像に対して金属アーチファクトをシミュレートしたアーチファクト合成画像を作成し，元画像とアーチファクト合成画像の差分から正解ラベル画像を作成した．

最初に，アーチファクト合成画像の作成手順を以下に示す．

STEP 1　元画像からシミュレーション用の金属ラベルとする歯の 3 次元領域を抽出し 2 値ボリュームラベルを作成する．

STEP 2　STEP1 で作成した 2 値ボリュームラベルと元画像を用いてアーチファクトをシミュレートし，アーチファクト合成画像を作成する．

STEP 3　アーチファクト合成画像内の金属ラベルを含むスライスのみ選択し，データセットに加える．

STEP 4　1 症例の元画像において 8 本の歯を選択し，1 本ずつを金属ラベルとして STEP 1 から STEP 3 を行う．

以上の処理を 29 症例に対して行い，それぞれにおいて 8 箇所の金属ラベルに基づいてアーチファクトをシミュレートしデータセットを作成した．STEP 1 における歯の 3 次元領域抽出は手作業で行った．STEP 2 では，従来の深層学習を用いたアーチファクト低減の研究 [10] で使用されたサイノグラム逆投影の手続き，及び，同じパラメータを用いて金属アーチファクトをシミュレートした．STEP 3 でデータセットに加えるスライスは各アーチファクト合成画像において 4〜29 枚からなる．アーチファクト合成画像内の各スライスの解像度は 512 × 512pixel，画素値は [−1000, 4000] である．アーチファクト合成画像から取得したスライス画像は計 2905 枚である．STEP 4 における 8 本の歯は，奥歯から 2 本とそれに近い奥歯から 2 本，さらに前歯から 2 本，最後に残りの歯から 2 本をそれぞれランダムに選択した．

次に正解ラベル画像の作成方法について説明する．まず STEP 2 の方法を用いて，金属ラベルを定義しないまま元画像の再構成を行った．再構成した元画像とアーチファクト合成画像の差分からラベルを作成することを試みた．アーチファクト合成画像はサイノグラム逆投影によって再構成した画像であるため，アーチファクト領域以外において元画像に対して差異が生じる可能性がある．この可能性を考慮し，元画像についてもサイノ

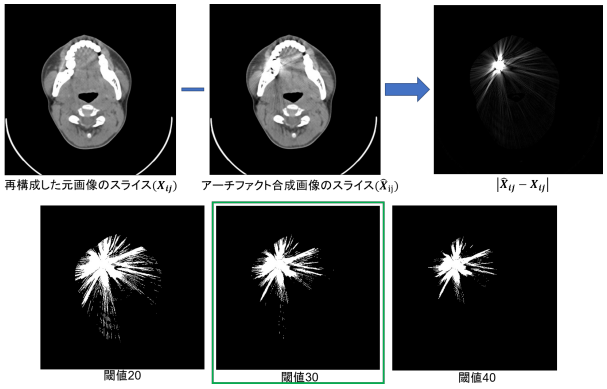

図 3　閾値ごとの正解ラベル画像

グラム逆投影による再構成を行った．

再構成した元画像における各画素を $\boldsymbol{X}_{ij}$ (i, j は画素の位置)，対応するアーチファクト合成画像の画素を $\hat{\boldsymbol{X}}_{ij}$ とする．本研究では，金属アーチファクトによる影響を受けたアーチファクト領域をアーチファクト合成画像と元画像の差分によって定義することとし，式 (7) に基づいてアーチファクト領域を示す正解ラベル画像のラベル値 $\boldsymbol{L}_{ij}$ を決定する．

$$\boldsymbol{L}_{ij} = \begin{cases} 1 & \text{if } |\hat{\boldsymbol{X}}_{ij} - \boldsymbol{X}_{ij}| \geq t \\ 0 & \text{if } |\hat{\boldsymbol{X}}_{ij} - \boldsymbol{X}_{ij}| < t \end{cases} \tag{7}$$

ここで t は閾値であり，$\hat{\boldsymbol{X}}_{ij}$ と $\boldsymbol{X}_{ij}$ の差異が一定値以上確認できる領域をアーチファクトによる影響を受けた領域であると定める．あるスライスについて元画像とアーチファクト合成画像，$|\hat{\boldsymbol{X}}_{ij} - \boldsymbol{X}_{ij}|$，さらに各閾値によって定められるアーチファクト領域を比較したものを図 3 に示す．可視化のため各正解ラベル画像が示すアーチファクト領域の画素値を 255 として強調した．閾値の大きさにより，定められるアーチファクト領域の範囲に差異が生じることを確認した．本実験では，複数のスライスにおいてアーチファクト合成画像と閾値ごとの正解ラベル画像を視覚的に比較し，閾値 30 で作成した正解ラベル画像を用いて実験を行った．

3.2　金属アーチファクト検出精度比較

3.2.1　実験環境・条件

本実験では，全 29 症例分のアーチファクト合成画像のうち 26 症例分を訓練データ，3 症例分をテストデータとして使用した．スライス枚数は訓練データ 2632 枚，テストデータ 273 枚である．Tensorflow-GPU で実装し、学習率 1×10^{-4} と重み減衰量 1×10^{-8} の RMSprop optimizer を使用してモデルをトレーニングした．データを入力する際には $[-1000, 4000]$ の画素値を $[0, 1]$ に正規化した．バッチサイズは 4，epoch 数は 60 とした．LambdaNetwork モデルと U-net のそれぞれにおいてスライス単位で学習し，アーチファクト検出を試みた．データ量が小規模であるためバリデーションデータとしてテストデータを使用した．損失関数には BCEWithLogitsLoss を用い，バリデーションロスが最小となったモデルで検出を行った．本実験は，CPU:Intel Core-i9，メモリ:64GB，GPU:NVIDIA TITAN RTX によって構成された計算機を用いた．

表 1　テストデータ 3 症例における金属ラベル位置ごとのアーチファクト合成スライス枚数

	位置1	位置2	位置3	位置4	位置5	位置6	位置7	位置8
症例A	11	11	16	12	15	12	10	9
症例B	10	12	10	10	13	12	11	12
症例C	9	10	9	10	12	11	13	13

3.2.2　評価指標

本実験では，比較評価を行うための評価指標としてダイス係数を用いる．ダイス係数による精度評価について説明するために以下を定義する．

- TP(True Positive) : アーチファクトを正しく検出した領域
- FP(False Positive) : 正解ラベルに存在しないアーチファクトを誤って検出した過検出領域
- FN(False Negative) : 正解ラベルに存在するアーチファクトを検出できなかった未検出領域

TP,FP,FN によってダイス係数は式 (8) で定義される．

$$Dice = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{8}$$

ダイス係数は 0 ~ 1 の範囲で与えられ，過検出領域や未検出領域が少ないほど 1 に近い値となる．スライス単位でダイス係数を算出し，アーチファクト領域の検出精度を評価する．

3.2.3　実験結果

表 1 にテストデータとして使用した 3 症例 (A,B,C) について，金属ラベル位置ごとのアーチファクト合成画像のスライス枚数を示す．アーチファクト検出を行ったスライスすべてに対してダイス係数を算出した．本節に示すダイス係数は小数点第四位を四捨五入した概数である．症例ごとのダイス係数を表 2，図 4 に示す．症例 A,B,C の項目に示す結果は，3 症例のテストデータ全てに対する結果を表す．表 2 と図 4 よりテストデータ 3 症例のいずれにおいても，LambdaNetwork モデルによるダイス係数が大きい傾向にあった．T 検定 ($p < 0.05$) により，症例 B，症例 C，症例 A,B,C において LambdaNetwork モデルと U-net の間に有意差が確認された．また，図 5 においていくつかのスライスに関して同一スライスにおける検出結果を手法間で比較した．上段・中段はそれぞれ LambdaNetwork モデル・U-net によるダイス係数が大きかった例であり，下段は両モデルによるダイスが小さかった例である．(b)(c) における緑色の領域はアーチファクトを正しく検出した領域であり，赤色の領域は過検出領域，白色の領域は未検出領域である．紫色の矢印が示す箇所では (b) に対して (c) の過検出領域が広く，橙色の矢印が示す箇所では，(b) でアーチファクトが検出されている一方で (c) ではほとんど検出されなかった．アーチファクトを正しく検出した領域が (c) よりも (b) で広く見られた箇所を黄色の矢印で，(b) よりも (c) の方で広く見られた箇所を青色の矢印で示す．下段のように他スライスと比較して，(a) に見られるアーチファクト領域に対する正解ラベル領域が小さいスライスではダイス係数が小さいにあった．

表 2　各症例に対するダイス係数 平均値 ± 標準偏差

	U-net	LambdaNetwork モデル
症例 A	0.751 ± 0.046	0.759 ± 0.044
症例 B	0.783 ± 0.033	0.801 ± 0.026*
症例 C	0.793 ± 0.022	0.812 ± 0.024*
症例 A,B,C	0.775 ± 0.040	0.790 ± 0.040*

*: $p < 0.05$（LambdaNetwork と U-net の 2 群検定）

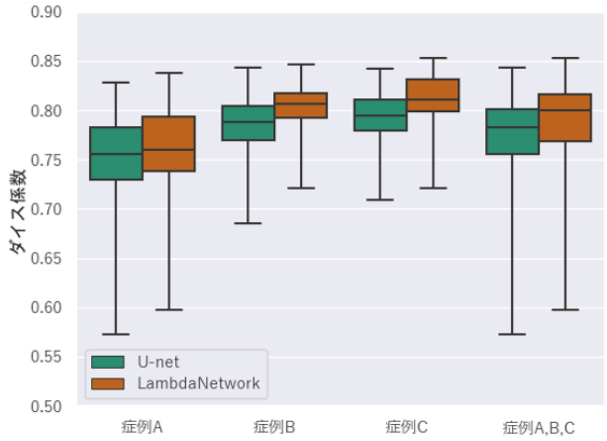

図 4　各症例に対するダイス係数

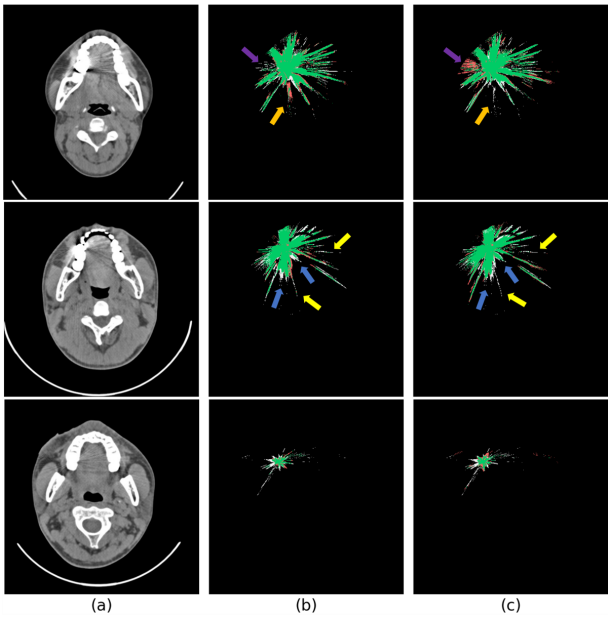

図 5　手法間比較 (閾値 30 の場合), (a) アーチファクト合成画像, (b) LambdaNetwork モデルによる検出結果, (c) U-net による検出結果

4. おわりに

本研究では，より広範囲にわたる画像間の関係性をもとに特徴抽出を行うことが広い画素領域をもつ金属アーチファクトを検出することに効果的であるという仮説から，LambdaNetwork を用いてアーチファクト検出を行った．U-net に LambdaNetwork を適用したモデルと U-net でアーチファクト検出精度を比較した．検出精度を定量評価するために，歯科金属を有しない 29 名の頭頸部 CT ボリュームに対してアーチファクトをシミュレートし，アーチファクト合成画像及び正解ラベル画像を作成した．2 つのモデルでアーチファクト領域を検出した結果に対してダイス係数を用いて評価し，閾値 30 として作成した正解ラベル画像に対する検出において LambdaNetwork の有効性を確認した．

今後の展望としては，詳細な検出領域おける有効性およびラベルの閾値が検出精度へ及ぼす影響の調査や，アーチファクト領域をより正確に示すラベル画像の作成方法の検討が必要である．

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．

文　献

[1] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, (2016).

[2] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 pp. 234–241, (2015).

[3] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, Deep convolutional networks do not classify based on global object shape, PLoS Computational Biology, Vol. 14, (2018).

[4] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, Proceedings of the 31st International Conference on Neural Information Processing Systems - NIPS 2017 pp. 6000–6010, (2017).

[5] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, Stand-alone self-attention in vision models, Advances in Neural Information Processing Systems, Vol. 32, (2019).

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv, art no.2010.11929, (2021).

[7] I. Bello, Lambdanetworks: Modeling long-range interactions without attention, arXiv, art no.2102.08602, (2021).

[8] M. Nakao, S. Aso, Y. Imai, N. Ueda, T. Hatanaka, M. Shiba, T. Kirita and T. Matsuda, "Automated Planning with Multivariate Shape Descriptors for Fibular Transfer in Mandibular Reconstruction", IEEE Trans. on Biomedical Engineering, Vol. 64, No.8, pp.1772-1785, (2017).

[9] H. Liao, W. Lin, J. Yuan, S. K. Zhou, and J. Luo, Artifact disentanglement network for unsupervised metal artifact reduction, in IEEE Transactions on Medical Imaging, vol. 39, no. 3, pp. 634-643, (2019).

[10] Y. Zhang and H. Yu, Convolutional Neural Network based Metal Artifact Reduction in X-ray Computed Tomography, IEEE Transactions on Medical Imaging, Vol. 37, pp. 1370–1381, (2018).

[11] T. Hase, M. Nakao, M. Nakamura, T. Matsuda, Improvement of Image Quality of Cone-beam CT Images by Three-dimensional Generative Adversarial Network, Proc. 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2843-2846, (2021)

[12] M. Nakao, K. Imanishi, N. Ueda, Y. Imai, T. Kirita, and T. Matsuda, Regularized Three-Dimensional Generative Adversarial Nets for Unsupervised Metal Artifact Reduction in Head and Neck CT Images, IEEE Access, Vol. 8, pp. 109453–109465, (2020).

[13] Y. Ou, Y. Yuan, X. Huang, K. K. Wong, J. Volpi, J. Z. Wang, and S. T. C. Wong, LambdaUNet: 2.5D Stroke Lesion Segmentation of Diffusion-Weighted MR Images, Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, pp. 731–741, (2021).