広島県立総合技術研究所東部工業技術センター研究報告　No.38（2025）
技術ノート

深層学習による画像の領域分割（第 3 報）

SegFormer

廣川　勝久

Image Segmentation using Deep Learning(Ⅲ)

SegFormer

HIROKAWA Katsuhisa

近年、自然言語モデル用に開発された Transformer 技術を画像認識へ適用した Vision Transformer が提案された。本報告では、Vision Transformer を基盤とした SegFormer[arXiv:2105.15203 (2021)]を実装し、一般画像および医用画像を用いてそのセグメンテーション性能を評価する。一般画像に対する実験では、完全な初期状態からの学習によってセグメンテーションが可能であることを確認する。医用画像を用いた評価では、撮影やコントラストにばらつきのある状況下でも、SegFormer が安定した性能を維持することを示す。

キーワード：SegFormer、Vision Transformer、location information leak、efficient self-Attention、semantic segmentation

1. 緒　　言

近年の 10 数年間にわたり、深層学習技術はめざましい進歩を遂げており、その応用範囲と性能は飛躍的に拡大している。従来は、多層パーセプトロン（**M**ulti-**L**ayer **P**erceptron**: MLP**）や畳み込みニューラルネットワーク（**C**onvolutional **N**eural **N**etwork**: CNN**）といったネットワーク構造が主流であったが、これに加えて近年では Transformer が新たな選択肢として注目を集めている[1)]。Transformer はもともと自然言語処理のために開発されたが、その優れた学習能力とスケーラビリティにより、深層学習モデルの大規模化におけるブレークスルーとなり、ChatGPT をはじめとする大規模言語モデルに広く採用されている[2)]。

加えて、Transformer の持つ Attention 機構は言語モデルのみならず他の分野への応用も試みられている。特に画像処理の分野においては、従来の CNN が持つ局所的な受容野の制約により、入力層に近い段階では画像全体の長距離依存関係を捉えることが困難であった。一方、Transformer の Attention 機構は、浅い層からでも広範な領域の情報を同時に処理できるため、画像全体の大局的な特徴を初期段階で捉えることが可能である。このような特性を活かし、近年では **Vi**sion **T**ransformer**: ViT** をはじめとする Transformer ベースのモデルが、画像認識、画像生成、セマンティックセグメンテーション（semantic segmentation）などに応用されている[3, 4)]。

我々のこれまでの報告では、セマンティックセグメンテーション技術が、航空写真における領域の抽出[5)]や、植物の葉に発生した病斑領域の検出に有効であることを示してきた[6)]。本報告では、撮影条件やコントラストのばらつきが大きく、領域の判別が困難となる工業用 X 線 CT 画像や医用画像への応用を想定し、セグメンテーションにおける境界領域の検出精度に優れるとされる SegFormer[7)]を実装し、一般画像および医用画像を用いてそのセマンティックセグメンテーション性能を評価する。SegFormer は ViT を基盤としながらも、階層型 Transformer エンコーダと軽量な MLP デコーダを組み合わせた、高精度かつ効率的なセマンティックセグメンテーション手法である。しかしその特性上、大量の学習データを必要とする可能性があるため、本稿ではまず一般画像を用いて完全な初期状態から学習を行い、セグメンテーション性能と学習収束性を評価する。続いて、医用画像を用いた評価では、撮影条件やコントラストにばらつきのある胸部 CT 画像データセットから解像度 512×512 画素の画像 15000 枚を用い、CT 画像に対する境界領域検出性能を評価する。

2. SegFormer

2.1 SegFormer の概要と基本構成

SegFormer は ViT における Transformer Encoder のアーキテクチャを基盤として、マルチスケールな特徴抽出を実現するセマンティックセグメンテーションモデルである。これまで報告した U-Net[8)]や PSPNet[9)]などと同様に、本モデルも Encoder–Decoder 型の二段構成を採用しており、Encoder においては解像度ごとに Attention 機構を用いた特徴抽出が行われる。その後、各解像度で得られた特徴マップは Decoder により統合され、セグメンテーション領域の再構成が行われる。このような構成により、異なる解像度における特徴マップの効果的な統合

2025. 7. 8　加工技術研究部　受理

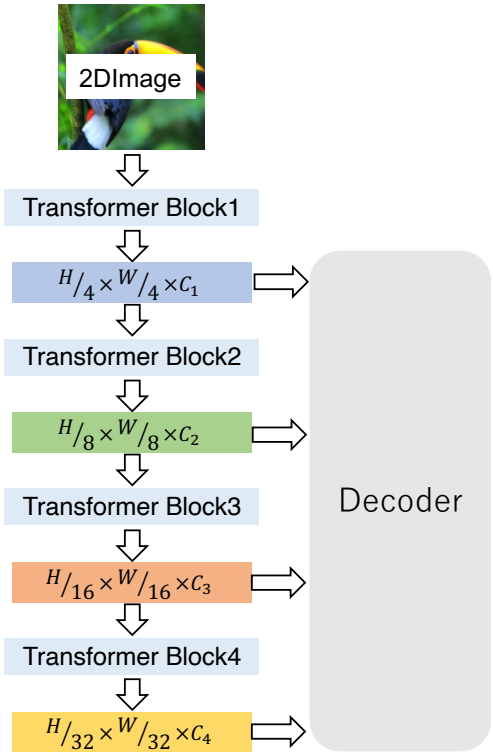


図 1　Encorder の構造

と、高精度な領域推定が可能となる。

2. 2 Encoder によるマルチスケール特徴抽出

SegFormer の Encoder の構造を図 1 に示す。Encoder は、入力画像から空間的特徴を抽出し、それらに位置情報を付加した特徴マップを生成することを目的とする。U-Net と同様、SegFormer の Encoder は多段に接続された Transformer Block により構成されており、各ステージにおいて画像は段階的に低解像度化される。その結果、複数スケールにおける特徴マップが生成され、後段の Decoder へと送られ、セマンティック領域の再構成に利用される。

2. 3 Transformer Block の構造

図 2 に Transformer Block の構造を示す。Transformer Block では、2 次元データが ViT と同様に 1 次元のトークン列に変換され、その後 Self-Attention および Mix-FFN による特徴抽出が行われる。また、Transformer Block においても、Attention 処理と Mix-FFN はスキップ接続されている。

ViT では 2 次元畳み込み層の stride と kernel のサイズを一致させることでパッチ間の重なりを排除していたが、SegFormer では kernel を stride よりも大きく設計することで、パッチ間の境界領域も含めた特徴抽出が可能となっている。また、ViT のような CLS トークンの追加は行わない。

Attention 処理は、ViT と同様、以下の式で定式化される。

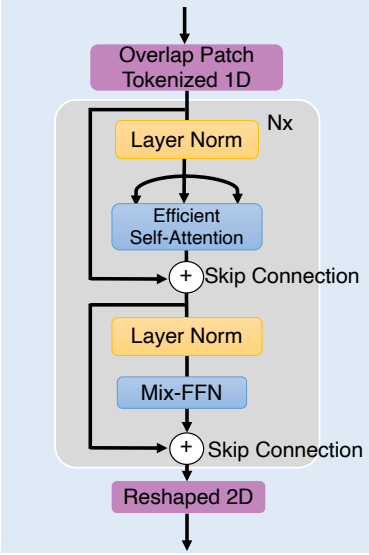


図 2　Transformer Block

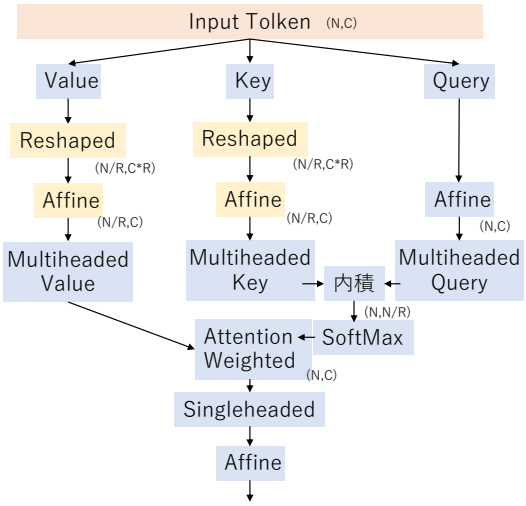


図 3　Attention 機構のフローチャート

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_{head}}}\right)V \qquad (1)$$

ここで、Key(K) と Value(V) に対しては、以下の処理により計算効率の向上が図られる。図 3 には高効率化された Attention 機構のフローチャートを示す。まず、Key(K) はトークン数Nを係数RによりN/Rに削減し、ベクトル次数は$C \times R$に拡張される。

$$\hat{K} = \mathrm{Reshape}\left(\frac{N}{R}, C \cdot R\right)(K) \qquad (2)$$

その後、学習可能な Affine 変換によりベクトル次数を元のC に圧縮する。

$$K = \mathrm{Linear}(C \cdot R, C)(\hat{K}) \tag{3}$$

Value(V) についても同様の圧縮処理がなされる。これによりトークン数は削減されるが、Attention 重み付け後の最終出力は元のトークン列と同じ次元に再構成される。

Attention により重み付けされた出力は、Mix-FFN に入力され、空間的特徴が次式に従って抽出される。

$$x_{out} = \mathrm{MLP}\left(\mathrm{GELU}\left(\mathrm{Conv}_{3\times3}\left(\mathrm{MLP}(x_{in})\right)\right)\right) + x_{in} \tag{4}$$

ViT は、全結合層 (MLP) のみの構造であったが、Mix-FFN では、3 × 3の 2 次元畳み込み層を中心に据え、その前後に全結合層 （MLP） を配置した構造を持つ。SegFormer では、構造的な理由から位置情報を埋め込まない。一方、畳み込み層の計算の過程にゼロパディングを実装することによって位置情報が漏れることが知られている (location information leak)。Mix-FFN では、この性質を活用することで、位置埋め込みを用いることなく位置情報の学習を実現している。

2. 4 Decoder によるセマンティックマップの再構成

Encoder により抽出されたスケールの異なる複数の特徴マップは Decoder に入力され、セマンティック領域として統合・再構成される。**図 4** に、Decoder が各スケールの特徴マップからセマンティック領域を生成する過程を示す。SegFormer の Decoder は、PSPNet と同様に、各スケールから得られた特徴マップを一括して処理するアプローチを採用している。まず、Encoder から出力される各スケールの特徴マップF_iに対してクラスター数C_iを同一次元のクラスター数Cに統一するための学習可能な線形変換が適用される。

$$\hat{F}_i = \mathrm{Linear}(C_i, C)(F_i), \forall_i \tag{5}$$

続いて、空間解像度を統一するため、各特徴マップの幅と高さを$W/4 \times W/4$にアップサンプリングする。

$$\hat{F}_i = \mathrm{Upsample}\left(\frac{W}{4} \times \frac{W}{4}\right)(\hat{F}_i), \forall_i \tag{6}$$

クラスター数および空間解像度が揃えられた各スケールの特徴マップは連結され、再度クラスター数をCに圧縮するための線形変換が適用される。

$$F = \mathrm{Linear}(4C, C)\left(\mathrm{Concat}(\hat{F}_i)\right), \forall_i \tag{7}$$

最終的に、各空間位置におけるクラス予測を得るために、クラス数N_{cls}に対応する線形変換が適用されたセグメンテーションマップが出力される。

$$M = \mathrm{Linear}(C, N_{cls})(F) \tag{8}$$

この出力に対して、対応するラベル画像との損失が計算され、モデル全体が学習される。

3. SegFormer の実装

3. 1 学習データ

SegFormer の性能評価には、Microsoft が公開する一般物体認識用データセットである **C**ommon **O**bjects in **Co**ntext: **COCO** データセット [10]および、胸部 CT 画像セグメンテーション（Chest CT Segmentation）データセット [11]を用いた。SegFormer は内部に Transformer アーキテクチャを備えており、大量の学習データを必要とする構造的特性を有する。そこで、まず COCO データセットを用いて、初期状態からの学習可能性および一般画像に

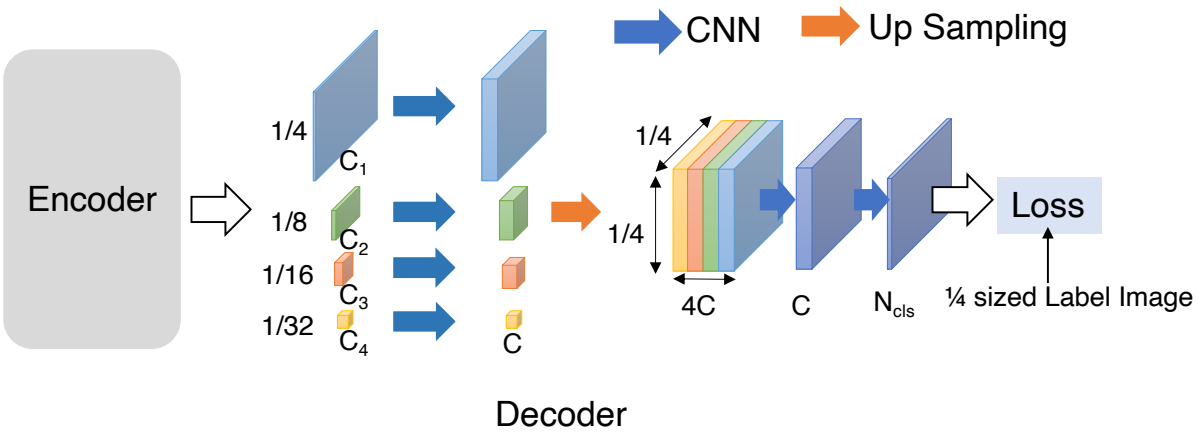


図 4　Decorder の構造

対するセグメンテーション性能の基礎的検証を実施した。一方、Chest CT Segmentation データセットを用いた評価では、X 線 CT 装置などの工業応用を想定し、撮影条件やコントラスト、対象物（人物等）が大きく異なる画像に対する SegFormer の汎化性能およびセグメンテーション能力の検証を行った。これら 2 種類の性質の異なるデータセットを通じて、SegFormer がセマンティックセグメンテーションモデルとして有する汎用性および応用可能性を総合的に評価した。

3. 2 SegFormer の具体的実装

SegFormer の Encoder は、多段に接続された Transformer Block により構成されており、入力画像に対して階層的に特徴抽出を行う構造を有する。各 Transformer Block では異なる空間解像度の特徴マップが生成され、それらは後段の Decoder に一括して入力される。

```
def forward(self, x):
    x1=self.block1(x)
    x2=self.block2(x1)
    x3=self.block3(x2)
    x4=self.block4(x3)
    return x1,x2,x3,x4
```

各 Transformer Block では、入力された 2 次元画像を 1 次元のトークン列に変換したのち、Self-Attention および Mix-FFN による処理を複数回適用し、最終的に再度 2 次元特徴マップへと再構成する。

```
def forward(self, x):
    x = self.input(x)
    x = self.dropout(x)
    for t in self.transformer:
        x = t(x)
    x = torch.permute(x,(0, 2, 1))
    h=np.sqrt(x.shape[2]).astype(int)
    x= x.reshape(x.shape[0], -1, h, h)
    return x
```

図 3 に示すような高効率化された Attention 機構は ViT における標準的な Attention 機構を一部変更することで、高効率な処理が実現されている。

```
def forward(self, x):
    x=self.lynorm(x)
    key   = x.reshape(x.shape[0], -1,
            x.shape[2]*self.ratio) #(N/R,C*R)
    Value = x.reshape(x.shape[0], -1,
            x.shape[2]*self.ratio) #(N/R,C*R)
    query = self.fcq(x)
    key   = self.fck(key)   #(C*R,C)
    Value = self.fcv(Value) #(C*R,C)
    query = query.reshape(query.shape[0],
             query.shape[1], self.headnum, -1)
    key   = key.reshape  (key.shape[0],
            key.shape[1],self.headnum, -1)
    Value = Value.reshape(Value.shape[0],
            Value.shape[1], self.headnum, -1)
    query = torch.permute(query,(0, 2, 1, 3))
    key   = torch.permute(key,  (0, 2, 3, 1))
    Value = torch.permute(Value,(0, 2, 1, 3))
    tmp = (query @ key)/np.sqrt(query.shape[3])
    tmp = self.softmax(tmp)
    tmp = self.dropout(tmp)
    tmp = torch.matmul(tmp,Value)
    tmp = torch.permute(tmp,(0, 2, 1, 3))
    tmp = tmp.reshape(x.shape[0],
              x.shape[1],x.shape[2])
    tmp = self.fc(tmp)
    x   = self.dropout(tmp)
    return x
```

Mix-FFN（式(4)）の実装では、従来の全結合層（MLP）をすべて kernel サイズが 1 の 2 次元畳み込み層に置き換えることで 2 次元データのままフィードフォワード処理が可能となっている。そのため、Attention 処理を通過した 1 次元トークン列は再び 2 次元空間へと変換され、以下のような畳み込み層により処理される。

```
self.fc1  = nn.Conv2d(dim,dim,kernel_size=1)
self.conv = nn.Conv2d(dim,dim*expand,
      kernel_size=K_size, padding=K_size//2)
self.fc2  = nn.Conv2d(dim*expand, dim,
                      kernel_size=1)
```

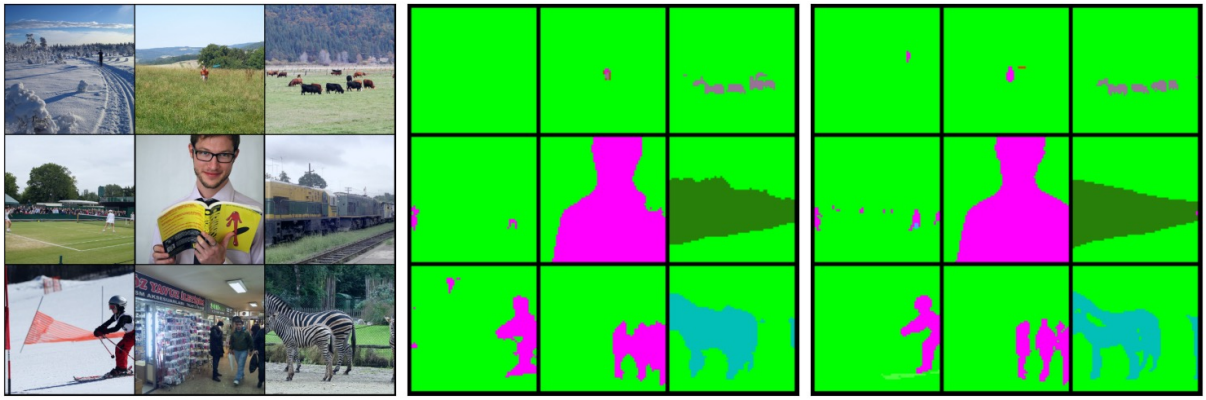
Decoder 部では、Encoder から出力された複数スケールの特徴マップを統合し、空間的整合性を保ちながら最終的なセグメンテーションマップを出力する。以下は、式(5)以降の処理を示した Decoder 部の実装例である。

```
def forward(self, x1,x2,x3,x4):
    x1 = self.block1(x1)
    x2 = self.block2(x2)
    x3 = self.block3(x3)
    x4 = self.block4(x4)
     x = torch.cat([x1,x2,x3,x4], dim=1)
     x = self.conv1(x)
     x = self.gn(x)
     x = self.gelu(x)
     x = self.conv2(x)
     return x
```

ここで、(5)式に示す全結合層（MLP）に対応した self.block も、以下に示すように kernel サイズが 1 の 2 次元畳み込み層に置き換えることで構成されており、(7)、(8)式に相当する処理も同様に実装される。

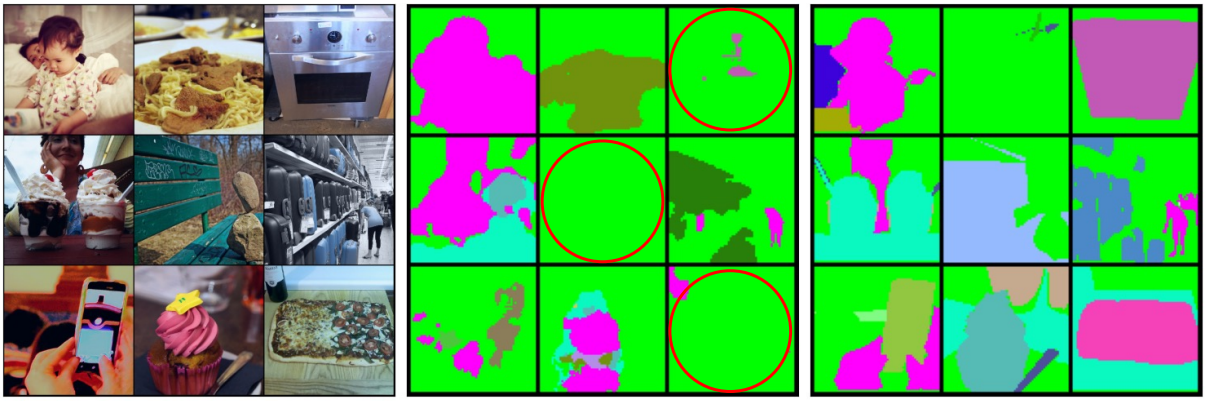
```
self.conv = nn.Conv2d(L1, C, kernel_size=1,
                      bias=False)
self.pool = nn.UpsamplingBilinear2d(size)
```

このように、SegFormer は Transformer の Attention 機構と、CNN による空間構造の保持を両立させたモジュ



(a) 入力画像　　(b) セグメンテーション結果　　(c) 正解ラベル画像

図 5　損失誤差が小さいセグメンテーション結果



(a) 入力画像　　(b) セグメンテーション結果　　(c) 正解ラベル画像

図 6　損失誤差が大きいセグメンテーション結果

ール構成により、セマンティックセグメンテーション処理を実現している。

4. SegFormer の性能評価

4.1 COCO データセットによる性能評価

SegFormer は内部に ViT Encoder に類似したエンコーダ構造を有しており、その構造的特性から大量の学習データを必要とすることが予想される。そこでまず、本実験では一般的な画像に対するセマンティックセグメンテーション性能評価として、COCO データセットを用いて完全な初期状態からの学習可能性と性能評価学習および検証を実施した。エンコーダ部分には、SegFormer の中で最小構成となる MiT-B0 を採用し、学習時には PSPNet により提案されたデータ拡張手法を適用した。学習には、COCO データセットからランダムに抽出した 256 画素×256 画素の画像 5 万枚を使用し、100 エポックの学習を行った。

学習後、モデルの学習可能性と性能を評価するため、学習に未使用の画像 50 枚を入力し、出力されたセグメンテーションマップを解析した。その実験結果として、誤差の小さい画像 9 枚と誤差の大きい画像 9 枚をそれぞれ**図 5** および**図 6** に示す。各図において、(a)は入力画像、(b)は SegFormer によって出力されたセグメンテーション結果、(c)は対応する正解ラベル画像を示す。本実験の結果、一般的な画像 5 万枚を用いた学習において、モデルが完全な初期状態から十分に収束することが確認された。また、誤差の小さい**図 5** からは、SegFormer が画像中に存在する非常に小さな物体まで正確に検出可能であるなど、高い空間分解能と認識性能を有することが示された。。一方、誤差の大きい**図 6** では、赤丸に示す画像全体に広がるような大きな物体に対して検出精度が低下する傾向が確認できた。これは、ViT と同様に Transformer Encoder 構造が持つ共通の課題であり、広範な空間からの特徴抽出における改善の余地を示した。

4.2 胸部 CT 画像データセットによる性能評価

工業用 X 線 CT 装置においては、CT 画像から内部構造の境界領域を人手により判定する作業が依然として存在している。これらの作業は、撮影条件や画像コントラスト、さらには作業者の主観的判断に依存するため、一定の誤差を伴うことが多い。そこで本実験では、X 線

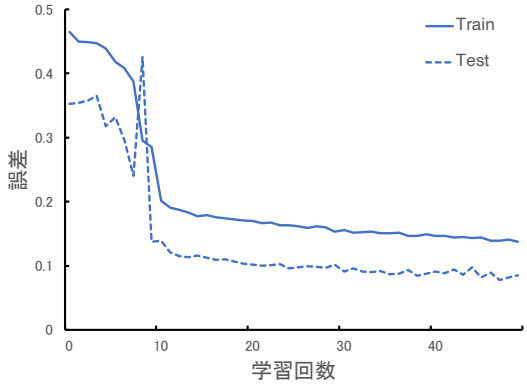
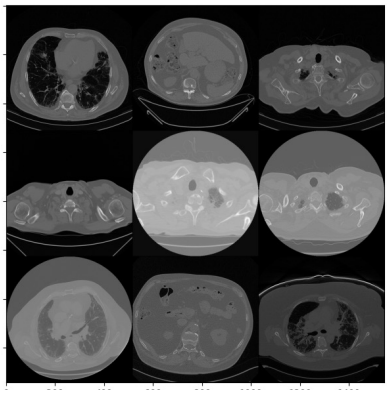


図 7　学習曲線

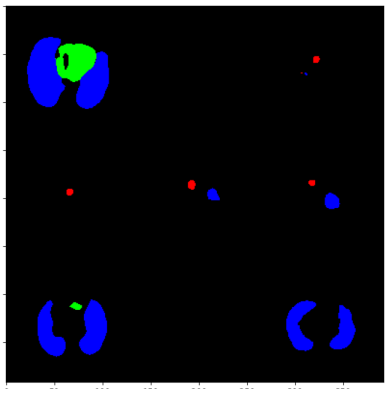
CT 装置などの工業応用を想定し、胸部 CT 画像データセットを用いて、撮影条件やコントラストが大きく異なる画像に対する SegFormer のセグメンテーション性能評価を行った。

実験では、胸部 CT 画像データセットから解像度 512×512 画素の画像 15000 枚を用い、SegFormer に対して 50 エポックの学習を行った。本実験でもモデルには、最小構成の MiT-B0 を採用した。学習過程における損失誤差の変化を**図 7** に示す。学習曲線より、胸部 CT 画像に対しては学習が急速に進行する領域が存在することが確認された。

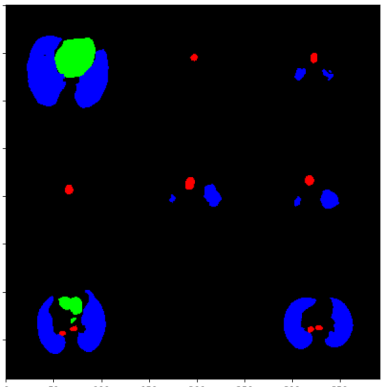
学習完了後、未学習のデータセットからランダムに 9 枚の画像を選び、SegFormer によるセグメンテーションを行った結果を**図 8** に示す。**図 8 (a)** は各入力画像を示しており、画像間で撮影条件にばらつきがあることが分かる。**図 8 (c)** は対応する正解ラベル画像であり、緑が心臓、青が肺、赤が気管の領域を示す。これと SegFormer から出力されたセグメンテーション結果（**図 8 (b)**）を比較したところ、心臓および肺については領域検出が行われていることが確認された。一方で、気管のようなサイズの小さい臓器に関しては、一部の画像において検出が困難であるケースが見られた。



(a) 入力画像



(b) セグメンテーション結果



(c) 正解ラベル画像

図 8　胸部 CT 画像のセグメンテーション結果

5. 結　　言

本報告では、セマンティックセグメンテーションモデルである SegFormer の実装を行い、一般画像および医用画像を対象とした性能評価を実施した。一般画像に対する実験では、完全な初期状態からの学習においても学習の収束が確認され、特に小さな物体に対しても高精度なセグメンテーションが可能であることが示された。また、医用画像を用いた評価においては、撮影条件やコントラストにばらつきがある画像に対しても、心臓や肺といった主要臓器の検出が良好に行われることが確認された。これらの結果より、SegFormer が持つセマンティックセグメンテーションモデルとしての有効性が実証された。さらに、本モデルは工業用 X 線 CT 装置などの工業検査における熟練作業者による目視確認の自動化や、多様な画像解析処理への応用が期待される。

文　　献

1) Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin: Attention is all you need, arXiv:**1706.03762** (2017).

2) Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec

Radford, Jeffrey Wu, Dario Amodei: Scaling Laws for Neural Language Models, arXiv:**2001.08361** (2020).

3） Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:**2010.11929** (2020).

4） 廣川 勝久：広島県立総合技術研究所東部工業技術センター研究報告，深層学習による画像認識（第1報），38（2025）.

5） 廣川 勝久，花房 龍男，中濱 久雄：広島県立総合技術研究所東部工業技術センター研究報告，深層学習による画像の領域分割（第1報），36（2023）.

6） 廣川 勝久，花房 龍男，中濱 久雄：広島県立総合技術研究所東部工業技術センター研究報告，深層学習による画像の領域分割（第 2 報），37（2024）.

7） Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, arXiv:**2105.15203** (2021)**.**

8） Olaf Ronneberger, Philipp Fischer, and Thomas Brox,: U-net: Convolutional networks for biomed-ical image segmentation, arXiv: **1505.04597** (2015).

9） Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia: Pyramid Scene Parsing Network, arXiv: **1612.01105** (2017).

10） Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár: Microsoft COCO: Common Objects in Context, arXiv:**1405.0312** (2014).

11） Chest CT Segmentation　Chest CT scans together with segmentation masks for lung, heart, and trachea, https://www.kaggle.com/datasets/polomarco/chest-ct-segmentation