長岡技術科学大学大学院
工学研究科修士論文

題　　目

Hierarchical Inter-Level Attention を用いた Transformer によるポリープ検出

指導教員
准教授　　杉田　泰則

著　　者
電気電子情報工学分野
信号処理応用研究室
20314689 NGUYEN HUU THO

令和 6 年 2 月 09 日

ABSTRACT

Polyp Segmentation using Hierarchical Inter-Level Attention with Transformer

Author: 20314689 NGUYEN HUU THO
Supervisor: Associate Professor YASUNORI SUGITA

Polyp segmentation is the process of precisely identifying polyps in the colon and analyzing their characteristics in detail using medical image analysis technology. Deep learning is widely used to take into account the diversity of polyp shapes and distributions, to reduce the burden on physicians, and to lead to early detection. Recent Transformers-based models have shown excellent results for polyp segmentation: the self-attention layer of Transformers-based models acts as a low-pass filter, effectively capturing long-range dependencies. The ColonFormer model achieves high detection accuracy with the same number of parameters as the CNN model. However, the ColonFormer encoder is a type of Hierarchical Visual Transformer (HVT) that often propagates feature information in one direction from low to high levels, which is not ideal. Local features are extracted at the Higher-level with low resolution, while global features are extracted at the Lower-level with higher resolution. Since there is important information at different stages, cross-propagation is considered necessary. In this paper, we propose a method using HILA to selectively extract important information by fusing Lower-level and Higher-level features with the aim of improving the accuracy of polyp detection. By applying HILA to a model using a Hierarchical Vision Transformer in semantic segmentation, we can better control which features are fused and which are not fused by Inter-level Attention. So we expect it to produce good results when combined with Colonformer on the polyp dataset.

In Experiment 1, the test results of applying HILA to one stage of ColonFormer confirmed that HILA works best when applying to stage 3 of ColonFormer for most of the datasets. When applied to two stages of ColonFormer, ColonFormer + HILA S23 and ColonFormer + HILA S34 were more accurate. These results suggest that the best results are obtained when applied HILA to stage 3 of the ColonFormer and combination between stage 3 and other stage.

In Experiment 2, we compared the accuracy of the conventional method and the ColonFormer + HILA S3 model by being the best effect when applied to stage 3 of the ColonFormer. Accuracy of ColonFormer B1 + HILA S3 was higher than that of the conventional method ColonFormer B1 for all data sets. The proposed method ColonFormer B1

+ HILA S3 outperformed ColonFormer B1 by 0.8% in DICE and 1.15% in IoU for the Kvasir dataset, and by 0.6% in DICE and 0.8% in IoU for the CVC-ClinicDB dataset. In the case of the Kvasir and CVC-ClinicDB datasets, the accuracy of the model that applied HILA to ColonFormer B1 stage 3 was higher than that of ColonFormer B2. The number of parameters of the proposed model ColonFormer B1 + HILA S3 was about 70% of that of the conventional method ColonFormer B2. For large data sets such as CVC-ColonDB or small polyps dataset like CVC-300 or high resolution data sets such as ETIS-LaribPolypDB, ColonFormer B2 was much more accurate. In other words, larger models are needed for more complex data sets.

目次

第 1 章

はじめに

1.1 研究背景

1.1.1 Semantic Segmentation

セマンティックセグメンテーションは、画像内の各ピクセルにラベルを割り当てるタスクである。オブジェクトは 1 つまたは複数の異なる物体である場合がある。これは特に医療画像や製造業などでよく使用される。このタスクは、各ピクセルにラベルとカテゴリ情報を関連付け、画像内の各ピクセルとラベルの集合を認識することによって実現される。特に自動運転技術では、車両、歩行者、交通標識、歩道、道路の特徴などの位置を特定し推定するために使用される。医療分野では、内視鏡や X 線の画像データからのスクリーニングや解析を通じて早期診断に役立つ。

1.1.2 Polyp Segmentation

大腸がん（Colorectal Cancer）の予防において、ポリープセグメンテーションが重要な役割を果たしている。大腸がんは通常、ポリープから進行して発症するため、ポリープの早期発見と適切な治療が必要である。

ポリープセグメンテーション (図 1.1) は、医療画像解析技術を用いて大腸内のポリープを精密に識別し、その特性を詳細に分析するプロセスである。ポリープの形状や分布の多様性を考慮し、医師の負担を軽減し、早期発見につながるために、深層学習が広く活用されている。近年、特にポリープセグメンテーションにおいて最も広く使用されている手法は、畳み込みニューラルネットワーク（CNN）に基づいている。ほとんどのセグメンテーションモデルは、エンコーダーとデコーダーを含む UNet[1] ベースのアーキテクチャ (Pranet[2]、Caranet[3] など) を採用しており、これらは通常畳み込み層から構築される。CNN はセグメンテーションタスクで優れたパフォーマンスを

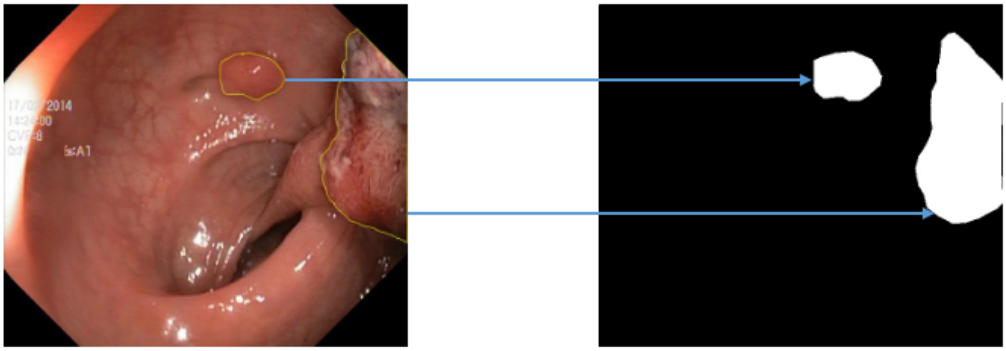


図 1.1: Polyp segmentation

発揮していますが、一定の制約がある。CNN は受容野が限定されているため、空間コンテキストとグローバル情報を無視し、ローカル情報しか取得できない。

一方に、最近の Transformers[4] ベースモデルの自己注意層はローパスフィルターとして機能し、効果的に長い範囲の依存関係を捉えることができる。Transformers ベースモデルの中に ColonFormer[5] モデルが高い検出精度を実現している。Transformer ベースモデルの失点はパラメーター数が大きいですが、ColonFormer では CNN モデルと同じパラメーター数であれば検出精度が ColonFormer の方が比較的によかった。なお、ColonFormer の Refinement Module によりポリープ境界の検出を改善できる。しかし、ColonFormer は特徴が順序的に生成すると最初ステージのローカル特徴がなくなる可能性が高い。なお、各レベルに異なる重要な情報がありますので、異なるステージの特徴を融合することが必要と考えられる。HILA [6] とは Hierarchical Inter-Level Attention の略である。セマンティックセグメンテーションにおける階層的ビージョントランスフォーマーを用いたモデルに HILA を適用することで、Inter-level Attention によってどの特徴を融合し、どの特徴を融合しないかをより適切に制御することができる。

1.2　研究目的

本論文では、ポリープを検出精度を向上することを目的として、Lower-level 特徴と Higher-level 特徴を融合することで、重要な情報を選択的に抽出する Hierarchical Inter-Level Attention を用いる手法を提案する。

1.3　本論文の構成

本論文の構成は次の通りである。第 1 章では、本論文の研究背景及び目的を述べた。第 2 章では、関連研究について述べる。第 3 章では、提案手法である ColonFomer に HILA を適用することについて述べる。第 4 章では、提案法と従来法の比較実験を行い、提案法の有効性を示す。第 5 章では、本論文を通してのまとめと今後の課題を述べる。

第 2 章

関連研究

2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNN) は、特にコンピュータビジョンの分野で広く使用される深層ニューラルネットワークのアーキテクチャである。深い層では、ますます抽象度が高くなる多層で特徴を抽出する。Lower-level 特徴は、高い解像度でグローバル特徴を示し、Higher-level 特徴は、低い解像度で豊富な意味情報を示す。

UNet [1] は、医療画像セグメンテーションのための先駆的な CNN アーキテクチャである。UNet は、エンコーダとデコーダのセットから構成される。エンコーダには、畳み込みおよびプーリングレイヤが含まれ、特徴を抽出する。デコーダには、アップサンプリング（またはデコンボリューション）レイヤと畳み込みレイヤが使用され、最終に画像の予測を行う。その後の研究では、スキップ接続を導入して UNet を改善し、多層畳み込みの重なりによる情報の損失を軽減した。ただし、低いレベルの情報を保持することで、性能を低下させるノイズ信号を生成する可能性がある。UNet の変種である UNet や DoubleUnet [7] は、ポリープセグメンテーションのデータセットで優れた結果を達成した。UNet は、異なる深さを持つ複数のネストされた UNet のセットとして構築され、一部のエンコーダを共有し、ディープスーパーバイズド学習を使用して共同で学習する。DoubleUNet は、2 つの UNet ブロックを重ね、ASPP [8] および SE [9] ブロックを使用して特徴表現能力を向上させる。

2.2 Vision Transformers

Vision Transformer (ViT) は、画像分類タスクに対する新しいアプローチである。論文 [10] で Dosovitskiy によって導入された。従来の畳み込みニューラル ネットワーク (CNN) とは異なり、ViT はトランスフォーマー アーキテクチャを利用して、画像

認識における最先端のパフォーマンスを実現する。この革新的なアプローチは、コンピューター ビジョンの分野に新たな可能性をもたらした。 ViT モデルは、画像を一連のトークンに分割し、セルフ アテンション メカニズムを適用して、画像カテゴリを予測するためのクラス トークンと一緒にそれらを処理する。

2.3　Hierarchical Vision Transformer

画像の解像度の固定で ViT は画像分類において優れた結果を示していますが、大きな画像のセマンティック セグメンテーションなどの高密度予測タスクに対しては計算効率が良くない。

なお、現在の ViT モデルのルーチンは、推論中に全長のパッチ シーケンスを維持することですが、これは冗長であり、階層表現がない。これらの問題により、PVT[11] や Swin[12] などのような階層型ビジョン トランスフォーマー (HVT) モデルが発展された。

階層型ビジョン トランスフォーマー（Hierarchical Vision Transformer、HVT）は、画像認識や関連するビジョンタスクにおいて、トランスフォーマーの力強い特性を組み合わせて高い性能を発揮するネットワークアーキテクチャである。HVT は階層的な特徴表現の取得に焦点を当てる。画像処理の初期段階では、低レベルの特徴（エッジ、テクスチャなど）が取り、これが徐々に高レベルの特徴（物体や概念）へと統合される。これにより、モデルは画像内の構造を理解し、より複雑なビジョンタスクに適応できる。

HVT の基本構造には、トランスフォーマーのアーキテクチャが採用される。トランスフォーマーは、セルフアテンションメカニズムを使用して長距離の依存関係を学習し、異なる位置の情報を組み合わせる能力に優れている。これにより、画像全体の文脈を理解し、重要な特徴を抽出できる。画像は小さなパッチに分割され、各パッチはトランスフォーマーの入力として処理される。

2.4　ColonFormer

2.4.1　全体モデル

Colonformer[5] は、エンコーダ-デコーダ構造と Refinement Module を組み合わせたポリープ セグメンテーションで使用されるモデルで、ポリープの境界をより効果的に検出するのに役立つ。全体構成は図 2.1 に示す。

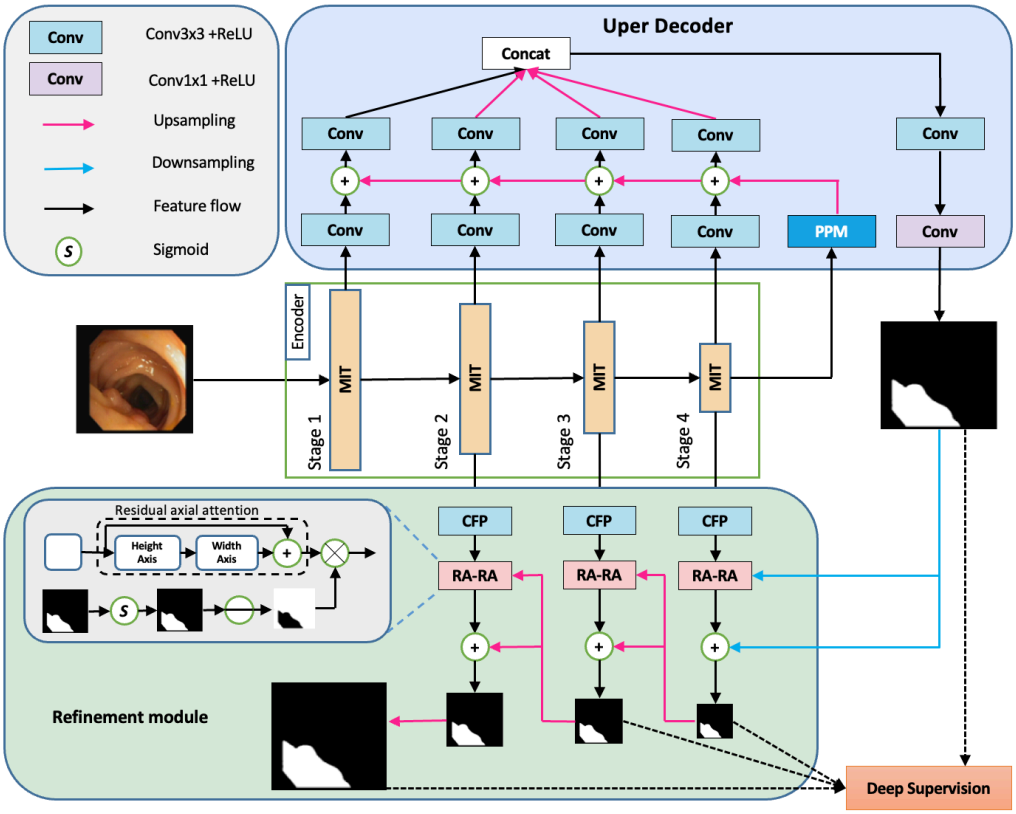


図 2.1: The overall architecture of ColonFormer ([5] より引用)

2.4.2　エンコーダ

Colonformer のモデルは、[13] で提案された Mix Transformer (MiT) をエンコーダのバックボーンとして使用される。 MiT は階層型ビジョン トランスフォーマーの一種であり、マルチスケール特徴マップを作成できる 。 $X \in \mathbb{R}^{H\times W\times C}$ を入力画像とし、 いくつかの Transformer ブロックの後 (図 2.2) に各ステージがそれぞれ $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$ を作成できる ($i \in \{1, 2, 3, 4\}$)。各ステージの間に Overlapped Patch Merging によりパッチごとのローカル特徴を保存できる。MiT ブロックには、Multi-head Self-Attention

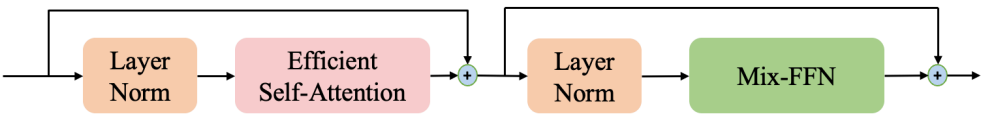


図 2.2: MIT block

(MHSA) Layer、Feed Forward Network (FFN)、および Feed Forward Network の 3 つの主要な部分が含まれる。 MHSA は、効果的なセルフアテンションに改善された。 次

に、セルフ アテンション レイヤーの計算の複雑さを軽減するために、キーの数が係数 R を減 る。 MiT には、MiT-B1 から MiT-B5 までのバージョンがあり、アーキテクチャは同じですがサイズが異なる。MiT ベージョンにそれぞれ対応して、ColonFormer B1、ColonFormer B2、ColonFormer B3、ColonFormer B4、ColonFormer B5 と名前を付ける。

2.4.3　デコーダ

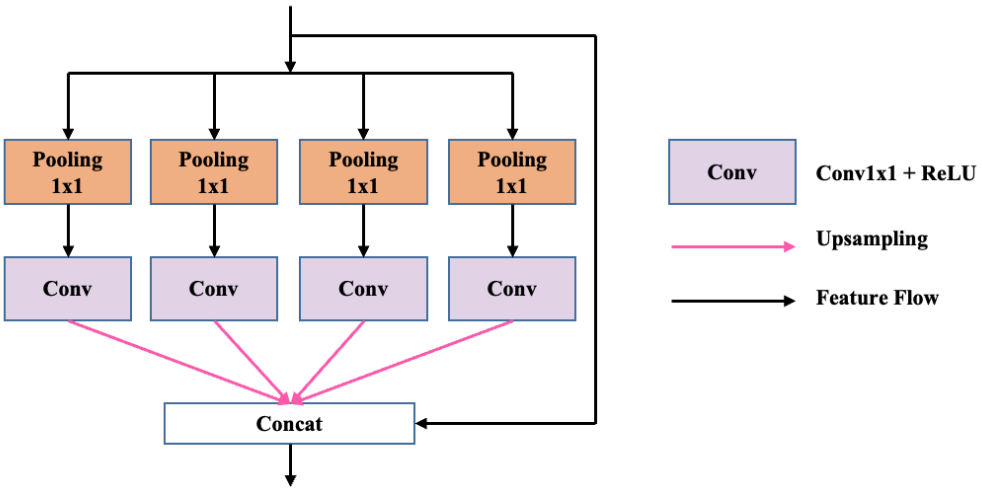


図 2.3: Pyramid Pooling Module (PPM)

エンコーダの最終ブロックから抽出された特徴マップは、デコーダ ブロックを通過する前に、まずピラミッド プーリング モジュール (PPM) [8] によって処理される。PPM は論文 [14] により紹介され、[8] によってさらに改善された。PPM は異なるサブ領域間のコンテキスト情報の損失をさらに減らすために、異なるスケールの情報を含み、異なるサブ領域間で異なる情報を含む階層的なグローバル事前分布を提案される。図 2.3 は、ピラミッド プーリング モジュールを詳細に示す。ColonFormers は、UPerNet [15] によりデコーダ アーキテクチャを使用し、これは UPer Decoder と呼ぶ。デコーダは、PPM によって生成された以前のグローバル マップを、MiT バックボーンによって生成されたマルチスケールの特徴マップと徐々に融合する。

2.4.4　Refinement Module

改良モジュールは、ポリープの境界を強調表示して、背景からポリープ領域を区別するパフォーマンスを向上させることを目的とする。そのモジュールには、Channel-wise Feature Pyramid (CFP) モジュール [16] (図 2.4) と、新しい残余軸に注意して強化され

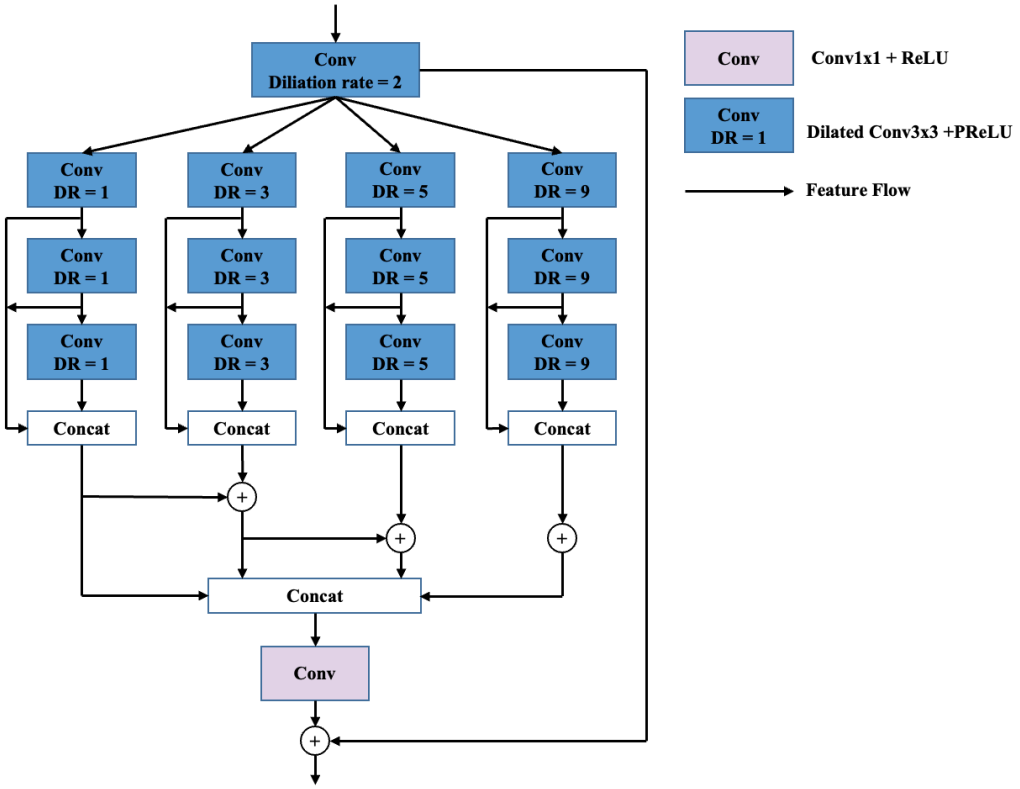


図 2.4: CFB block

た新しい逆軸アテンション ブロックが含まれ、段階的に調整する。アテンションモジュール Reserse Attention[17] と Axial attention [18] が含まれる。 [2] で導入された Parallel Reverse Attention network architecture では、ポリープの大まかな位置特徴のみを含む特徴マップが最も深い層から取得される。逆アテンションを使用すると、深い層で予測されたポリープ領域が省略され、代わりにより深い層からアップサンプリングされて、より正確で完全な特徴マップが作成される。 RA-RA モジュールに入る前に、各ステージ (s1、s2、s3) の特徴は、チャネルごとの特徴ピラミッド (CFP) で導入された技術であるチャネルごとの特徴ピラミッド CFP ブロックを通過する。 CaraNet [3] より、 マルチスケール ビューでエンコーダから特徴を抽出するために使用される。

2.4.5　問題点

ColonFormer のエンコーダ部では，ステージ 1 からステージ 4 まで特徴量が順番的に生成される。一般に，ローカル特徴を描写する情報は，Lower-level 特徴に現れる。一方，グローバル特徴は Higher-level 特徴に現れる。具体的にローカル特徴はポリープのエッジ、小さな特徴などが含める。グローバル特徴は観察範囲が広いので、ポリープの全体構成を見ることができる。特徴を順序的に生成すると最初ステージのローカ

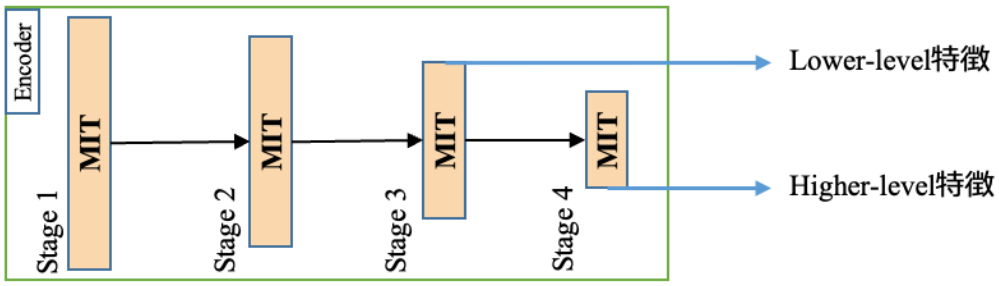


図 2.5: ColonFormer's encoder

ル特徴がなくなる可能性が高い。なお、各レベルに異なる重要な情報がありますので、両方を保存することが必要であり、両方を保存することが必要と考えられる。以上のことで、図 2.6 に表示している偽陰性と偽陽性の検出の原因になる可能性が高い。

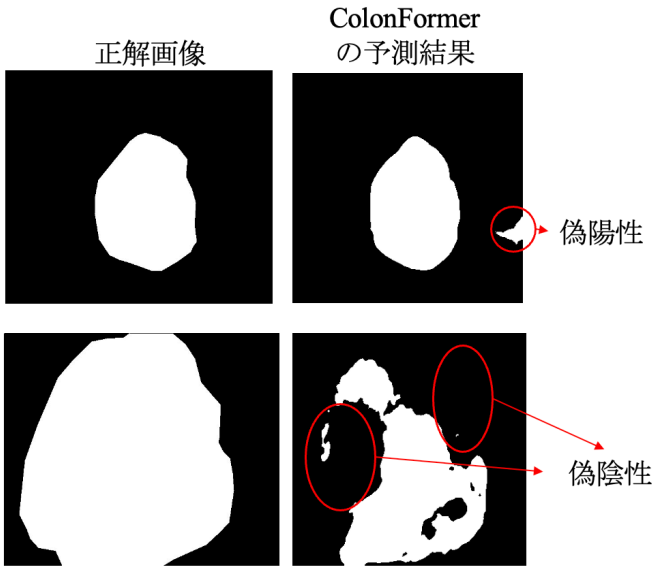


図 2.6: False negatives and false positives in polyp segmentation

従来法でも DenseNet [20]、UNet [33] などのモデルは、異なるスケール特徴を融合することで特徴マップを改善した。その中に Skip-connection という技術がありますが、特徴を融合するプロセスにコントロールがなくて、特徴がそのまま融合された。必要ない情報を伝番すると悪い影響を与える可能性があるので、特徴を融合する時に必要ない情報を抑える方法が必要と考えられる。

2.5　Hierarchical Inter-Level Attention

2.5.1　Inter-level Connections

HVT エンコーダのパフォーマンスを向上させるために最も人気のある方法の 1 つは、情報の融合である。従来の HVT では、特徴はパッチのマージングプロセスを通じて順次生成され、この初期のステップを超えて異なる階層レベル間で直接の接続を確立しない。階層間の接続の導入は、トランスフォーマーエンコーディングステップ中に異なるレベル間で反復的な相互作用を促進することで、この制限を克服することを目的とする。DenseNet [19] や最近では D3Net [20] など、いくつかのエンコーダは、バックボーンエンコーダに密な接続を持ち、機能を下位レベルから上位レベルに向けて一方向の融合機能を伝播させる。HRNet [21] は、単一の融合モジュールを介して両方向の畳み込み接続を持つエンコーダを導入する。

別の研究の分野では、画像から部分-全体の階層を抽出することに焦点を当てており、視覚データ内の関係と構造を理解することを目指する。この研究の先駆者として Capsules と GLOM があり、それぞれ階層的な特徴の改善に向けた新しい手法を紹介する。

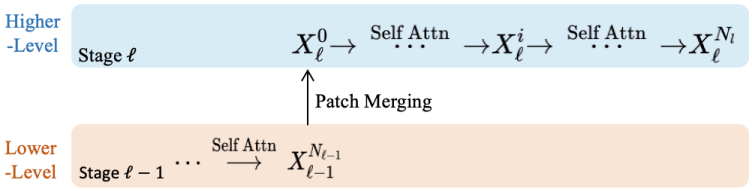
Capsules は、[22] などの論文で紹介されるように、反復的な EM アルゴリズムを使用する。このアルゴリズムは、高レベルのカプセルと低レベルの特徴の間の割り当てを解決することで、視覚階層内の複雑な関係を明確化するのに役立つ。

GLOM [23] が提案した手法であり、ボトムアップとトップダウンの相互作用の組み合わせを使用して、すべてのレベルで特徴を反復的に洗練する。GLOM の手法は、階層的な特徴の洗練に根ざしており、包括的な方法で部分-全体の階層を明らかにしようとする。

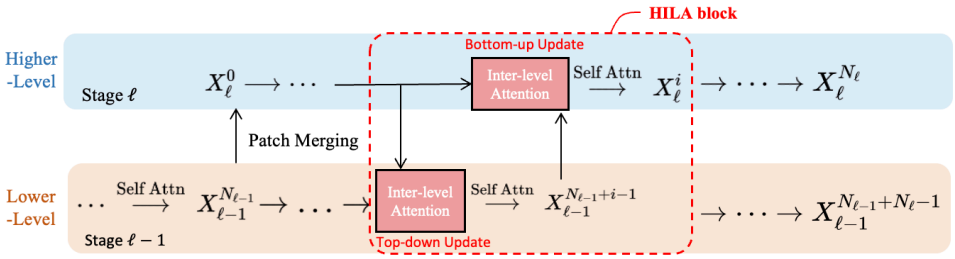
Visual Parser [24] は、この分野での別の取り組みであり、特徴を部分レベルと全体レベルに分割する。これらの異なる特徴は互いに反復的に更新されますが、更新プロセスは階層内の同じレベルに限定される。

2.5.2　Hierarchical Inter-Level Attention

HILA[6]（Hierarchical Image Level Attention）は、Transformer ブロックにステージ ℓ と前のステージ $\ell - 1$ の間にコネクションを作る。HILA のアイデアは Higher-layer の特徴を Lower-Level の特徴を融合することである。なお、Inter-Level Attention ブロックにより Self-attention により Lower-level 特徴と Higher-level 特徴の関係重みを計算し、その重みにより特徴融合をコントロールできる。これらの主なアップデート



(a) Standard Transformer Stage



(b) Standard Transformer Stage with HILA

図 2.7: HILA Block. (a) shows the standard hierarchical vision transformer block. (b) HILA block with Bottom-Up and Top-Down Update.

の詳細な説明は、論文の 2.5.3 節と 2.5.4 に記載される。

図 2.7(b) は、標準の HVT ステージに HILA を追加したものを示す。ベースの HVT アーキテクチャに対して、HILA を異なるステージに適用する。

X_ℓ^0 を Higher-level 特徴として、$X_{\ell-1}^{N_{\ell-1}+i}$ は Lower-level 特徴になる。最初の HILA ブロック ($i = 1$) は Top-down Update をスキップし、Bottom-up Update のみを適用する。(図 4b 参考) HILA がステージ ℓ に適用される場合、現在のステージ ℓ を Higher-level、前のステージ $\ell - 1$ を Lower-Level と呼ぶ。$i (i \in 0, \cdots, N_\ell)$ ループにおいて、更新した Higher-level 特徴は X_ℓ^i 、更新した Lower-level 特徴は $X_{\ell-1}^{N_{i-1}+i}$ である。Higher-level の最後に生成された特徴を i Top-down Update ブロックを通して、最後に $X_{\ell-1}^{N_{\ell-1}+N_\ell-1}$ の特徴が生成される。

2.5.3　Bottom-Up Update

このプロセスは Bottom-Up Inter-Level Attention と Self-attention Layer が含める。Bottom-Up Inter-Level Attention (図 2.8(b)) により Higher-level 特徴と Lower-level 特徴の融合を行う。図 2.8(a) により Higher-level 特徴マップ $X_{\ell,\{h,w\}}^i$ の中にローカルパッチ P_{hw} を選び、そのパッチに対応する 16 つ Lower-level 特徴が含める。Inter-Level Attention プロセスにより各 Lower-level と Higher-level のローカルパッチ P_{hw} に意味的に関係が大きほど Higher-layer に伝播する際に影響力が大きくなる。その時に、特

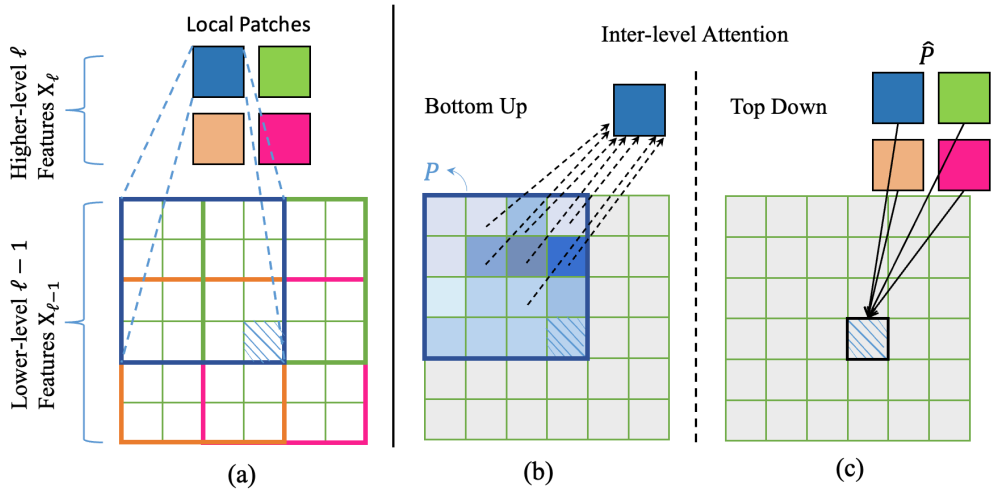


図 2.8: Top-Down and Bottom-Up Inter-Level Attention.

徴が選択的にアップデートされ、必要ない情報の伝番を抑えることができる。

Inter-Level Attention では、[4] の通り標準形式の内積アテンションを使用する。プロセスの入力は Higher-level 特徴 $X^i_{\ell,\{h,w\}} \in \mathbb{R}^{1\times d_\ell}$ と Lower-level 特徴 $X^{N_{\ell-1}+i}_{\ell-1,P_{hw}} \in \mathbb{R}^{16\times d_{\ell-1}}$ がある。これらをクエリ $Q = q\left(X^i_\ell\right)$、キー $k\left(X^{N_{\ell-1}+i}_{\ell-1,P}\right)$、値 $V = v\left(X^{N_{\ell-1}+i}_{\ell-1,P}\right)$ に投影する。ここで、q、k、v は異なるピクセル間で共有される完全接続レイヤーである。次に、以下の演算を使用してアテンションを計算する。

$$\text{attn}\left(X^i_\ell, X^{N_{\ell-1}+i}_{\ell-1,P}\right) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_\ell}} + B\right)V \tag{2.1}$$

相対位置エンコーディング B は、Swin Transformer [12] に続いて追加される。対応する Bottom-Up Inter-Level Attention の更新は次のようになる。

$$X^{i\prime}_\ell = X^i_\ell + \text{attn}\left(X^i_\ell, X^{N_{\ell-1}+i}_{\ell-1,P}\right) \tag{2.2}$$

$$X^{i\prime\prime}_\ell = \alpha X^{i\prime}_\ell + \beta\text{FFN}\left(X^{i\prime}_\ell\right) \tag{2.3}$$

Bottom-Up Update の中間出力を i' と i'' とする。α と β は、前のイテレーションからどれだけの情報を持ち越すかを制御するハイパーパラメータである。

Self-Attention レイヤーについて、Inter-Level Attention を使用して Higher-level の特徴を更新した後、同じレベルの特徴間でこの情報を伝播させる。これは、ベースの HVT モデルの Transformer ブロックから直接 Self-Attention レイヤーを利用することによって実現される。

$$X^{i+1}_\ell = \text{SelfAttn}\left(X^{i\prime\prime}_\ell\right) \tag{2.4}$$

2.5.4 Top-Down Update

Top-Down Update は、Higher-level の特徴で Lower-level の特徴を更新する。このプロセスも Inter-level Attention と Self-Attention Layer が含める。

Top-Down Inter-Level Attention Layer では、各 Lower-level の特徴 $X^{N_{\ell-1}+i}_{\ell-1,\{h,w\}} \in \mathbb{R}^{1\times d_{\ell-1}}$ は、最大 4 つの上位レベルの特徴 $X^{i}_{\ell,\hat{P}*hw} \in \mathbb{R}^{n\times d*\ell}$ のローカルパッチエリアによってカバーされる。ここで、$\hat{P}_{hw}$ は Higher-level の特徴の位置を示す ($n \in \{1,2,3,4\}$)。$n = 4$ の場合 Lower-level の特徴の大部分は 4 つの Higher-level の特徴にカバーされる。$n = 1$ の場合に 1 つ Higher-level の特徴がカバされる (画像の境界など)。

Top-Down Inter-Level Attention Layer は、図 2.8(c) に示す。このプロセスでは Higher-level の特徴が互いに競合し、意味的に類似した上位レベルの機能が Lower-level 特徴の更新に使用される。Top-Down Inter-Level Attention の更新は次のように計算する。

$$X^{N_{\ell-1}+i\prime}_{\ell-1} = X^{N_{\ell-1}+i}_{\ell-1} + \text{attn}\left(X^{N_{\ell-1}+i}_{\ell-1}, X^{i}_{\ell,\hat{P}}\right) \tag{2.5}$$

$$X^{N_{\ell-1}+i\prime\prime}_{\ell-1} = \alpha X^{N_{\ell-1}+i\prime}_{\ell-1} + \beta\,\text{FFN}\left(X^{N_{\ell-1}+i\prime}_{\ell-1}\right) \tag{2.6}$$

Self-Attention Layer については Top-Down Update の Self-Attention Layer と同じように Self-Attention Layer Update を行う。

$$X^{N_{\ell-1}+i+1}_{\ell-1} = \text{SelfAttn}\left(X^{N_{\ell-1}+i\prime\prime}_{\ell-1}\right) \tag{2.7}$$

第 3 章

提案手法

3.1 ColonFormer+HILA の提案

HILA を ColonFormer のエンコーダのステージに適用する。エンコーダの構成に論文の 2.4 節 のように Inter-level Attention Self-Attention を導入する。

3.2 提案モデル

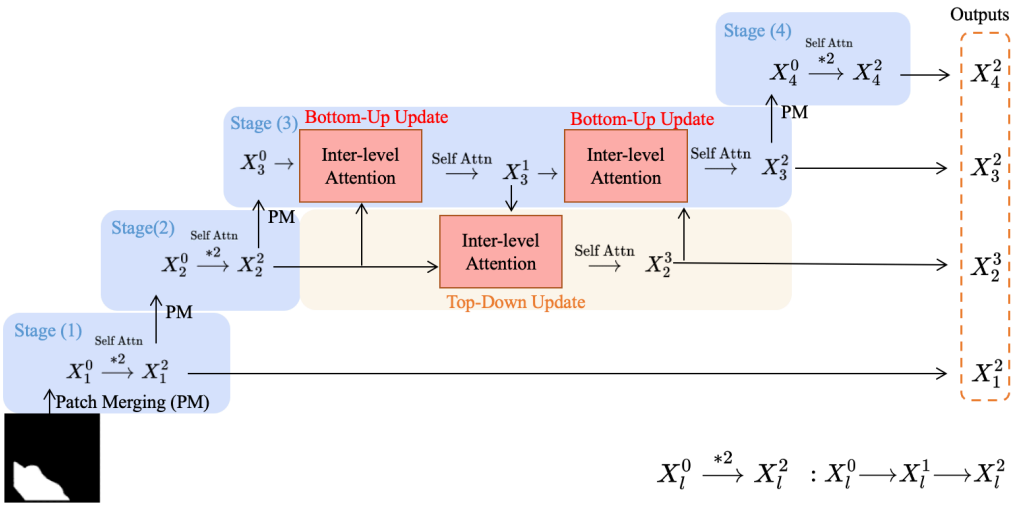


図 3.1: ColonFormer's Encoder with HILA applied to Stage 3 (HILA S3)

図 3.1 では ColonFormer's Encoder のステージ 3 に HILA を適用した構成である。HILA をステージ 3 に導入する場合、ステージ 3 に Inter-level Attention モジュールを入れて、ステージ 2 に最後に生成された特徴マップを用いて Bottom-up Update を行い、ステージ 3 の特徴マップを更新する。ステージ 2 に Inter-level Attention モジュールを入れて Top-down Update を行い、ステージ 3 の更新した特徴マップで更新する。

ステージ 3、4 に適用する場合もステージ 3 に適用する時のプロセスと同じである。(図 3.2 を参考)

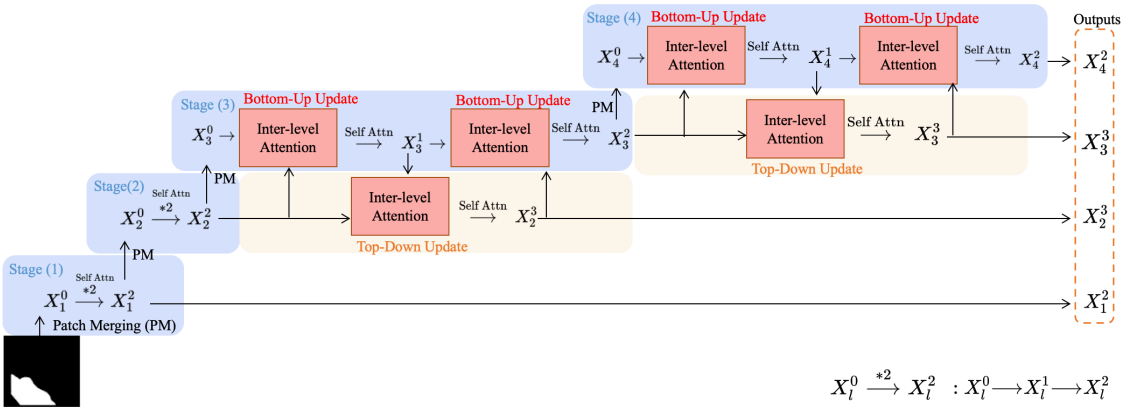


図 3.2: ColonFormer's Encoder with HILA applied to Stage 3, Stage 4 (HILA S34)

3.3　モデルの設定

表 3.1: Definition of symbols

	Parameters	Definition
ステージ	K_ℓ	パッチマージコンボリューションカーネルサイズ
	S_ℓ	パッチマージコンボリューションストライドサイズ
	d_ℓ	ステージのチャンネルサイズ
	N_ℓ	ステージ内のブロックの数
	H_ℓ	アテンションに使用されるヘッドの数
	E_ℓ	フィードフォワード層の拡大率
	R_ℓ	空間縮小時の縮小率注意
HILA	α_ℓ, β_ℓ	情報伝播定数
	s_ℓ	HILA が適用されるストライド
	p_ℓ	Higher-level 特徴と Lower-level 特徴の間のローカルパッチサイズ

添え字 ℓ は、ハイパーパラメータの段階を示し、以下の表に使用したパラメーターは表 3.1 に現れる。従来法の各ステージの詳細設計は表 3.2 に示す。提案法のネット

ワークの各ステージの詳細情報は表 3.3 と表 3.4 に示す。H と W はそれぞれ入力画像の高さ、幅である。

HILA の Top-Down Update の Self-Attention Layer は、元のバックボーンの Self-Attention ブロックと同じハイパーパラメータを使用する。デコードヘッドについては、すべてのバックボーンモデルの元のモデルと同じパラメータを使用する。

3.4　モデルの学習

提案法では、ColonFormer と同じように Weighted Focal Loss と Weighted IoU Loss という損失関数として使用する。

Weighted Focal Loss は以下の式で定義される。

$$\mathcal{L}_{wfocal} = -\frac{\sum_{i=1}^{H}\sum_{j=1}^{W}(1+\lambda\beta_{ij})\alpha\left(1-q_{ij}\right)^{\gamma}\log\left(q_{ij}\right)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(1+\lambda\beta_{ij}\right)} \tag{3.1}$$

α, γ は調整可能なハイパーパラメータである。

Weighted IoU Loss は以下の式で定義される。

$$\mathcal{L}_{wiou} = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(g_{ij} * p_{ij}\right) * \left(1+\lambda\beta_{ij}\right)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(g_{ij} + p_{ij} - g_{ij} * p_{ij}\right) * \left(1+\lambda\beta_{ij}\right)} \tag{3.2}$$

λ は、重要度の重み β_{ij} の影響を調整するためのハイパーパラメータである。

学習プロセスにピクセル間によりも重要であるピクセルがある可能性がある。このピクセル (i, j) の重要性を重み β_{ij} で表する。β_{ij} は以下の式に定義される。

$$\beta_{ij} = \left| \frac{\sum_{m,n\in\mathcal{N}_{ij}} gmn}{\left|\mathcal{N}_{ij}\right|} - g_{ij} \right| \tag{3.3}$$

p_{ij} はポリープクラスに属するピクセル (i, j) の予測確率である。p_{ij} は以下の式に定義される。

$$q_{ij} = \begin{cases} p_{ij}, & \text{if } g_{ij} = 1 \\ 1 - p_{ij}, & \text{otherwise} \end{cases} \tag{3.4}$$

ColonFormer の損失関数は、式 3.1 と 3.2 平均によって計算される。

$$\mathcal{L}_{\text{total}} = \frac{\mathcal{L}_{\text{wfocal}} + \mathcal{L}_{\text{wiou}}}{2} \tag{3.5}$$

最適化手法として Adam を用いて学習される。学習率は $1e-4$ に 設定する。

表 3.2: ColonFormer's architecture specifications

Stage	Output Size	Component	ColonFormer B1	ColonFormer B2
Stage 1	$\frac{H}{4} \times \frac{W}{4}$	Patch Merging	$K_1 = 7$ $S_1 = 4$ $d_1 = 64$	$K_1 = 7$ $S_1 = 4$ $d_1 = 64$
		Self-Attention-Layer	$R_1 = 8$ $H_1 = 1$ $E_1 = 4$ $N_1 = 2$	$R_1 = 8$ $H_1 = 1$ $E_1 = 4$ $N_1 = 2$
Stage 2	$\frac{H}{8} \times \frac{W}{8}$	Patch Merging	$K_2 = 3$ $S_2 = 2$ $d_2 = 128$	$K_2 = 3$ $S_2 = 2$ $d_2 = 128$
		Self-Attention-Layer	$R_2 = 4$ $H_2 = 2$ $E_2 = 4$ $N_2 = 2$	$R_2 = 4$ $H_2 = 2$ $E_2 = 4$ $N_2 = 2$
Stage 3	$\frac{H}{16} \times \frac{W}{16}$	Patch Merging	$K_3 = 3$ $S_3 = 2$ $d_3 = 320$	$K_3 = 3$ $S_3 = 2$ $d_3 = 320$
		Self-Attention-Layer	$R_3 = 2$ $H_3 = 5$ $E_3 = 4$ $N_3 = 2$	$R_3 = 2$ $H_3 = 5$ $E_3 = 4$ $N_3 = 6$
Stage 4	$\frac{H}{32} \times \frac{W}{32}$	Patch Merging	$K_4 = 3$ $S_4 = 2$ $d_4 = 512$	$K_4 = 3$ $S_4 = 2$ $d_4 = 512$
		Self-Attention-Layer	$R_4 = 1$ $H_4 = 8$ $E_4 = 4$ $N_4 = 2$	$R_4 = 1$ $H_4 = 8$ $E_4 = 4$ $N_4 = 2$

表 3.3: ColonFormer + HILA architecture specifications

Stage	Output Size	Component	ColonFormer	
			B1 + HILA S3	B1 + HILA S234
Stage 1	$\frac{H}{4} \times \frac{W}{4}$	Patch Merging	$K_1 = 7$ $S_1 = 4$ $d_1 = 64$	$K_1 = 7$ $S_1 = 4$ $d_1 = 64$
		Self-Attention-Layer	$R_1 = 8$ $H_1 = 1$ $E_1 = 4$ $N_1 = 2$	$R_1 = 8$ $H_1 = 1$ $E_1 = 4$ $N_1 = 2$
Stage 2	$\frac{H}{8} \times \frac{W}{8}$	Patch Merging	$K_2 = 3$ $S_2 = 2$ $d_2 = 128$	$K_2 = 3$ $S_2 = 2$ $d_2 = 128$
		Inter-level Attention Layer	N/A	$\alpha_2, \beta_2 = 0.5, 0.5$ $s_2 = 1$ $p_2 = 4$ $H_2 = 1$ $E_2 = 4$
		Self-Attention-Layer	$R_2 = 4$ $H_2 = 2$ $E_2 = 4$ $N_2 = 2$	$R_2 = 4$ $H_2 = 2$ $E_2 = 4$ $N_2 = 2$

表 3.4: ColonFormer + HILA architecture specifications

Stage	Output Size	Component	ColonFormer	
			B1 + HILA S3	B1 + HILA S234
Stage 3	$\frac{H}{16} \times \frac{W}{16}$	Patch Merging	$K_3 = 3$ $S_3 = 2$ $d_3 = 320$	$K_3 = 3$ $S_3 = 2$ $d_3 = 320$
		Inter-level Attention Layer	$\alpha_3, \beta_3 = 0.5, 0.5$ $s_3 = 1$ $p_3 = 4$ $H_3 = 1$ $E_3 = 4$	$\alpha_3, \beta_3 = 0.5, 0.5$ $s_3 = 1$ $p_3 = 4$ $H_3 = 1$ $E_3 = 4$
		Self-Attention-Layer	$R_3 = 2$ $H_3 = 5$ $E_3 = 4$ $N_3 = 2$	$R_3 = 2$ $H_3 = 5$ $E_3 = 4$ $N_3 = 2$
Stage 4	$\frac{H}{32} \times \frac{W}{32}$	Patch Merging	$K_4 = 3$ $S_4 = 2$ $d_4 = 512$	$K_4 = 3$ $S_4 = 2$ $d_4 = 512$
		Inter-level Attention Layer	N/A	$\alpha_4, \beta_4 = 0.5, 0.5$ $s_4 = 1$ $p_4 = 4$ $H_4 = 1$ $E_4 = 4$
		Self-Attention-Layer	$R_4 = 1$ $H_4 = 8$ $E_4 = 4$ $N_4 = 2$	$R_4 = 1$ $H_4 = 8$ $E_4 = 4$ $N_4 = 2$

第 4 章

実験

4.1 実験条件

4.1.1 データセット

モデルは Kvasir と CVC-ClinicDB のトレーニングセット (1450 枚) で学習され、Kvasir と CVC-Clinic のテストセットと他のデータセットで評価される。

Kvasir[25]：ノルウェーの Vestre Viken Health Trust (VV) で内視鏡機器を使用して収集されたデータセットである。画像は、VV およびノルウェーのがん登録所の経験豊富な胃腸科医によって注意深く注釈付けおよび検証されている。このデータセットは、720 × 576 から 1920 × 1072 ピクセルまでのさまざまな解像度の 1000 枚の画像で構成されている。

CVC-ClinicDB[26]：大腸内視鏡のビデオからフレームを抽出したデータベースである。このデータセットは、31 の内視鏡シーケンスから抽出された 384 × 288 ピクセルの 612 枚の画像で構成されている。このデータセットは、MICCAI 2015 の自動ポリープ検出チャレンジのトレーニング段階で使用された。

CVC-ColonDB[27]：Machine Vision Group（MVG）によって提供された CVC-ColonDB は、15 本の短い大腸内視鏡動画から取得した解像度が 574 × 500 ピクセルの 380 枚の画像で構成されている。

CVC-300[28]：CVC-300 は、36 人の患者から取得された 44 のビデオシーケンスから得られた 60 枚の画像で構成される、より広範なデータセットでのテストセットである。Endoscene と呼ばれるより大規模なデータセットの一部である。

ETIS-LaribPolypDB[29]：コロノスコピー手術からの 196 枚の高解像度画像（1226 × 996）のコレクションである。

表 4.1: Experiment Conditions

Test-set	Number of images
Kvasir	100
CVC-ClinicDB	62
CVC-300	60
CVC-ColonDB	380
ETIS-LaribPolypDB	62
Total	664

4.1.2　評価指標

図 4.1 により偽陽性、偽陰性、真陰性、真陽性の判別方を現れる。

FP: False Positive, FN: False Negative, TN: True Negative, TP: True Positive

図 4.1: Difference between FP, FN, TN, TP

IoU（Intersection over Union：名前が示すように、IoU（Intersection over Union）は、2 つのセット間の類似性を表す指数である。和集合 (Union) に対して共通集合 (Intersection) の割合で計算する。2 つのセットが A と B であると仮定すると、IoU は以下の式で計算される。

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{4.1}$$

IoU は 0 から 1 の範囲を取り、1 に近いほど類似性が高いことを示する。ただし、A と B の両方が空集合である場合、IoU は 1 となる。

Precision：予測された陽性の中で実際に陽性である割合を示す指標です。つまり、モデルが陽性と予測したもののうち、実際に陽性であるものの割合を計算します。

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

適合率は、偽陽性を最小限に抑えることが重要な場合に特に重要です。例えば、がんの検出などの医療診断では、誤って陽性と判定することがあると、不必要な治療や心理的な負担を引き起こす可能性があります。そのため、適合率は高いほど、モデルの正確性が高いと言えます。

Recall : 実際の陽性のうち、モデルが正しく陽性と予測した割合を示す指標である。つまり、全体の陽性サンプルのうち、モデルが正しく陽性と予測できた割合を計算する。

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

再現率は、モデルが陽性サンプルを見逃すことを最小限に抑えることが重要な場合に特に重要である。例えば、がんの検出などの医療診断では、陽性サンプルを見逃すことがあると、適切な治療や早期の介入の機会を逃す可能性がある。そのため、再現率は高いほど、モデルの網羅性が高いと言える。

Dice : Precision と Recall の調和平均になる。

$$\text{Dice} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4.4}$$

4.1.3　実験条件

表 4.2: Experiment Common Conditions

Train Dataset	90% of Kvasir and ClinicDB (1450 images)
Test Dataset	Kvasir, ClinicDB, CVC-300, CVC-ColonDB, ETIS-LaribPolypDB
Evaluation function	Dice, IoU, Precision, Recall
Batch size	8
Learning Rate	1e − 4
Optimization Method	ADAM

実験共通条件は表 4.2 に示す。学習データは Kvasir と ClinicDB の 90% である。残りはテストデータとして使用する。表 4.3 により実験 1 では HILA を異なるステージに適するの調査を行う。ImageNet[30] データセットで事前に ColonFormer の Pretrained Weight は HILA 重みがない、再トレーニングする際にランドムを選ぶ。HILA を適用の範囲は一つステージと二つステージである。

表 4.4 により実験 2 では HILA があるモデルと従来法の結果を比較し、HILA の効果を確認する。実験 2 では ImageNet データセットで ColonFormer+HILA の Pretrained Weight（HILA 重みがある）を使ってトレーニングを続く。HILA を適用の範囲は一つステージと三つステージである。

表 4.3: Experiment 1 Conditions

Model	ColonFormer + HILA
Mix-transformer Version	B1
Pretrained Model	ColonFormer の Pretrained Weight
Apply HILA for	Stage 2, Stage 3, Stage 4 Stage 2-3, Stage 3-4, Stage 2-4

表 4.4: Experiment 2 Conditions

Model	ColonFormer	ColonFormer + HILA
Mix-transformer Version	B1, B2	B1
Pretrained Model	ColonFormer の Pretrained Weight	ColonFormer + HILA の Pretrained Weight
Apply HILA for	なし	Stage 3 Stage 2-3-4

4.2　実験結果

4.2.1　実験 1

HILA を ColonFormer の一つステージに適用したの結果は表 4.5 に示す。テスト結果により、ほぼのデータセットに対して ColonFormer のステージ 3 に HILA を適用すると一番効果があると確認できた。表 4.6 により ColonFormer の二つステージに適用する場合、ColonFormer + HILA S23 と ColonFormer + HILA S34 の精度が高かった。

これらの結果は、ColonFormer のステージ 3 に HILA を適用し、ステージ 3 と他のステージの組み合わせに HILA を適用した場合に最良の結果が得られると考えられる。

表 4.5: Compare Results when applied HILA to one stage

Dataset	ColonFormer + HILA	Dice	IoU	Precision	Recall
Kvasir	**S2**	**0.922**	**0.869**	0.941	**0.920**
	S3	0.918	0.866	0.940	0.914
	S4	0.919	0.867	**0.944**	0.914
CVC-ClinicDB	S2	0.924	0.875	0.926	0.930
	S3	**0.930**	**0.880**	0.926	**0.948**
	S4	0.927	0.878	**0.929**	0.943
CVC-300	S2	0.887	0.815	0.846	**0.958**
	S3	**0.892**	**0.824**	**0.856**	0.956
	S4	0.875	0.802	0.832	**0.958**
CVC-ColonDB	S2	0.792	0.709	0.809	**0.827**
	S3	**0.810**	**0.729**	**0.850**	0.817
	S4	0.794	0.712	0.815	0.821
ETIS-LaribPolypDB	S2	0.753	0.673	0.733	0.845
	S3	0.756	0.677	0.742	0.840
	S4	**0.762**	**0.686**	**0.750**	**0.858**

4.2.2　実験 2

テスト結果によって、すべてのデータセットに従来法と比較して ColonFormer B1 のモデルに HILA を適用したモデルの検出精度の方は高かった。提案法では、表 4.7 と表 4.8 により ColonFormer に対して Kvasir データセットの場合、DICE で 0.8%、IoU で 1.15%、CVC-ClinicDB データセットの場合、DICE で 0.6% 、IoU で 0.8% 上回る結果が得られた。なお、Kvasir と CVC-ClinicDB データセットの場合 ColonFormer B1 のステージ 3 に HILA を適用したモデルの精度が ColonFormer B2 より高かった。表 4.12 により、提案モデル ColonFormer B1+ HILA S3 のパラメーター数は従来法の 70% 程度であることが確認できた。CVC-ColonDB のような大規模データセットや、ETIS-LaribPolypDB のような高解像度データセットでは、ColonFomrer B2 の方がはるかに精度が高い。つまり、より複雑なデータセットに対しては、より大きなモデル

表 4.6: Compare Results when applied HILA to two stage

Dataset	ColonFormer + HILA	Dice	IoU	Precision	Recall
Kvasir	**S23**	**0.920**	**0.871**	**0.941**	**0.919**
	S24	0.919	0.865	0.940	0.916
	S34	0.914	0.865	0.937	0.913
CVC-ClinicDB	**S23**	**0.936**	**0.888**	**0.930**	**0.950**
	S24	0.930	0.884	0.925	0.945
	S34	0.932	0.885	0.929	**0.950**
CVC-300	S23	0.866	0.792	0.823	**0.957**
	S24	0.893	0.825	0.861	0.952
	S34	**0.897**	**0.830**	**0.875**	0.944
CVC-ColonDB	S23	0.793	0.711	0.816	0.821
	S24	0.802	0.722	0.829	**0.825**
	S34	**0.807**	**0.726**	**0.843**	0.820
ETIS-LaribPolypDB	**S23**	**0.766**	**0.691**	**0.746**	**0.840**
	S24	0.744	0.668	0.734	0.822
	S34	0.739	0.667	0.730	0.798

が必要だと考えられる。

図 4.2 によって、従来法の結果にはポリープでは ないところをポリープを検出されたが、HILA を適用すると偽陽性の検 出がなくなりました。図 4.3 により、 従来法の結果はポリ ープであるところですが、ポリープではないと判断した。HILA を適用したら偽陰性の検出が改善できました。図 4.4 により、複数ポリープがあると精度がある程度に上がったが、スコアが低かった。図 4.5 に より、HILA があるモデルの結果が悪くなったこともある。以上のことで、小さいポリープ対して、改善方法が必要と言える。なお、提案法に対して精度が悪くなった時もあるので、モデルのハイパーパラメータを調査することが必要と考えられる。

表 4.7: Results of a model without HILA and with HILA in Kvasir Test-set

	ColonFormer B1			ColonFormer B2
モデル	従来法 HILA なし	HILA-S3	HILA-S234	従来法 HILA なし
DICE	0.917	0.925	0.919	0.923
IoU	0.865	0.875	0.871	0.870
Precision	0.940	0.947	0.946	0.941
Recall	0.913	0.919	0.914	0.919

表 4.8: Results of a model without HILA and with HILA in CVC-ClinicDB Test-set

	ColonFormer B1			ColonFormer B2
モデル	従来法 HILA なし	HILA-S3	HILA-S234	従来法 HILA なし
DICE	0.924	0.930	0.938	0.927
IoU	0.875	0.882	0.888	0.878
Precision	0.924	0.918	0.931	0.924
Recall	0.935	0.959	0.952	0.947

表 4.9: Results of a model without HILA and with HILA in CVC-300 Test-set

	ColonFormer B1			ColonFormer B2
忘デル	従来法 HILA なし	HILA-S3	HILA-S234	従来法 HILA なし
DICE	0.878	0.879	0.872	0.891
IoU	0.808	0.807	0.802	0.821
Precision	0.843	0.842	0.835	0.854
Recall	0.952	0.955	0.955	0.956

表 4.10: Results of a model without HILA and with HILA in CVC-ColonDB Test-set

	ColonFormer B1			ColonFormer B2
忘デル	従来法 HILA なし	HILA-S3	HILA-S234	従来法 HILA なし
DICE	0.794	0.802	0.800	0.810
IoU	0.714	0.723	0.718	0.729
Precision	0.810	0.843	0.805	0.811
Recall	0.840	0.822	0.847	0.868

表 4.11: Results of a model without HILA and with HILA in ETIS-LaribPolypDB Test-set

	ColonFormer B1			ColonFormer B2
忘デル	従来法 HILA なし	HILA-S3	HILA-S234	従来法 HILA なし
DICE	0.758	0.769	0.761	0.787
IoU	0.681	0.700	0.682	0.710
Precision	0.744	0.746	0.729	0.746
Recall	0.853	0.849	0.869	0.910

表 4.12: Number of parameters of different model

MODEL	Parameters (M)
ColonFormer B1	22.00
ColonFormer B1+ HILA S2	22.56
ColonFormer B1+ HILA S3	23.69
ColonFormer B1+ HILA S4	27.70
ColonFormer B1+ HILA S23	28.25
ColonFormer B1+ HILA S24	29.38
ColonFormer B1+ HILA S34	29.91
ColonFormer B1+ HILA S234	31.22
ColonFormer B2	33.04

(a) Grouth Truth

(b)従来法
ColonFormer B1

Dice : 0.941
IoU :0.887
Precision : 0.937
Recall : 0.893

(c) 提案法
ColonFormer B1
+HILA S3
Dice : 0.950
IoU : 0.904
Precision : 0.986
Recall : 0.917

図 4.2: Predicted Image Example 1 (Kvasir)

(a) Grouth Truth

(b) 従来法
ColonFormer B1

Dice : 0.588
IoU : 0.416
Precision : 0.869
Recall : 0.444

(c) 提案法
ColonFormer B1
+HILA S3
Dice : 0.932
IoU : 0.872
Precision : 0.989
Recall : 0.880

図 4.3: Predicted Image Example 2 (Kvasir)

(a) Grouth Truth

(c) 従来法
ColonFormer B1

Dice : 0.455
IoU : 0.295
Precision : 0.683
Recall : 0.341

(b) 提案法
ColonFormer B1
+HILA S3
Dice : 0.614
IoU : 0.444
Precision : 0.858
Recall : 0.479

図 4.4: Predicted Image Example 3 (Kvasir)

(a) Grouth Truth

(b) 従来法
ColonFormer B1

Dice : 0.949
IoU : 0.903
Precision : 0.983
Recall : 0.917

(c) 提案法
ColonFormer B1
+HILA S3
Dice : 0.935
IoU : 0.878
Precision : 0.900
Recall : 0.974

図 4.5: Predicted Image Example 4 (Kvasir)

第 5 章

おわりに

5.1　まとめ

本論文では、ポリープセグメンテーションの精度を向上することを目的に、Higher-level 特徴と Lower-level 特徴を選択的に融合する ColonFormer と HILA の組み合わせる手法を提案した。第 1 章では、ポリープセグメンテーションに関する手法、本論文の研究目的について述べた。第 2 章では、従来法 ColonFormer について述べた。なお、HILA という特徴融合の技術も述べた。第 3 章では提案法とモデルの設定について説明した。第 4 章では、提案法の有効性を確認するために、同一の学習データを用いて ColonFomer + HILA と提案法の評価実験を行った。

5.2　今後の課題

本論文では、従来法である ColonFormer に基づき Hierarchical Inter-Level Attention を用いる手法を提案した。しかし、提案法のモデルはまだ最適化されていないことで、ハイパーパラメータを調整することが必要である。なお、複雑なデータセットに対しては、ColonFormer B2 ような大きなモデルに HILA の適用を調査することが必要だと考えられる。

謝辞

本研究を進めるにあたり、ご指導・ご助言を頂いた杉田泰則准教授に厚くお礼申し上げます。また、論文の審査において多くのご指示を頂きました、本学電気系岩橋政宏教授、圓道知博教授ならびに原川良介准教授に厚く御礼申し上げます。最後に、本研究に関して多くの指摘をくださいました信号処理応用研究の皆様に深く感謝いたします。

参考文献

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[2] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International conference on medical image computing and computer-assisted intervention*, pp. 263–273, Springer, 2020.

[3] A. Lou, S. Guan, and M. Loew, "Caranet: context axial reverse attention network for segmentation of small medical objects," *Journal of Medical Imaging*, vol. 10, Feb. 2023.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[5] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, and V. S. Dinh, "Colonformer: An efficient transformer based method for colon polyp segmentation," *IEEE Access*, vol. 10, pp. 80575–80586, 2022.

[6] G. Leung, J. Gao, X. Zeng, and S. Fidler, "Hila: Improving semantic segmentation in transformers using hierarchical inter-level attention," *arXiv:2207.02126*, 2022.

[7] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 558–564, 2020.

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CoRR*, vol. abs/1709.01507, 2017.

[10] 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," *ICLR*, 2021.

[11] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvtv2: Improved baselines with pyramid vision transformer," *CoRR*, vol. abs/2106.13797, 2021.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021.

[13] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems (NeurIPS)*, 2021.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *CoRR*, vol. abs/1406.4729, 2014.

[15] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," *CoRR*, vol. abs/1807.10221, 2018.

[16] A. Lou and M. H. Loew, "Cfpnet: Channel-wise feature pyramid for real-time semantic segmentation," *CoRR*, vol. abs/2103.12212, 2021.

[17] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," *CoRR*, vol. abs/1807.09940, 2018.

[18] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," *CoRR*, vol. abs/1912.12180, 2019.

[19] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016.

[20] N. Takahashi and Y. Mitsufuji, "D3net: Densely connected multidilated densenet for music source separation," *Audio and Speech Processing (eess.AS)*, 2021.

[21] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *CoRR*, vol. abs/1904.04514, 2019.

[22] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *CoRR*, vol. abs/1710.09829, 2017.

[23] G. E. Hinton, "How to represent part-whole hierarchies in a neural network," *CoRR*, vol. abs/2102.12627, 2021.

[24] S. Sun, X. Yue, S. Bai, and P. H. S. Torr, "Visual parser: Representing part-whole hierarchies with transformers," *CoRR*, vol. abs/2107.05790, 2021.

[25] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and

H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pp. 451–462, 2020.

[26] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, vol. 43, pp. 99—111, July 2015.

[27] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. C. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *CoRR*, vol. abs/1612.00799, 2016.

[28] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2016.

[29] J. S. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Towards embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.