AUTHOR GUIDELINES FOR ICASSP 2020 PROCEEDINGS MANUSCRIPTS

Makoto Takamatsu

Tokyo Denki University

ABSTRACT

The abstract should appear at the top of the left-hand column of text, about 0.5 inch (12 mm) below the title area and no more than 3.125 inches (80 mm) in length. Leave a 0.5 inch (12 mm) space between the end of the abstract and the beginning of the main text. The abstract should contain about 100 to 150 words, and should be identical to the abstract text submitted electronically along with the paper cover sheet. All manuscripts must be in English, printed in black ink.

***Index Terms*—** Multi frequency extraction, semantic segmentation.

1. INTRODUCTION

大腸がん（CRC）は，世界中でがんによる死亡の 3 番目の原因である [1]．CRC は，最初は良性の腺腫性ポリープ（腺腫）から発生するが，時間が経つにつれて，それらのいくつかは悪性になる．CRC の生存率は，検出された段階に依存し，I 期の 95% の高い割合から，治療をせずに放置をした IV および V 期には 35% より低い割合に低下し致命的になる []．したがって，早期に病変やポリープを検出することが CRC の予防には不可欠である []．現在，CRC 関連の死亡率を減らすための標準的なアプローチは，ポリープを探すために定期的なスクリーニングを実施することであり，大腸内視鏡検査が最適なスクリーニングツールであり，臨床医はポリープを探すために腸壁を視覚的に検査する．しかしながら，内視鏡医が限られた時間内で大量の画像を読影しなければならず非常に負担が大きく，内視鏡医の誤検出率は約 25%[] と誤検出が起こりやすい．近年になり，ポリープのセグメンテーションは，慎重に検査するべき潜在的な病変で覆われている領域を定義するため，臨床医が除去する可能性を高めることやポリープ検出のような弱い探索はポリープの見落としにつながる可能性があるため，内腔のセグメンテーションは探索された結腸壁の程度に関連する品質指標を確立するために使用できることから，内腔内のポリープにセグメンテーションをする研究が行われている．

大腸内視鏡ビデオは，キャプチャ中，腸壁が常に動き揺れているため，観察を妨げる強いボケのあるビデオをキャプチャすることがよくあり，長時間，病変に焦点を合わせ続けることが困難である．そのため，内視鏡画像には多くの弱いエッジ，テクスチャのない領域，ぼやけた領域をもつ画像が存在し，診断のパフォーマンスを大幅に妨げる．今までは，エッジ情報が豊富なシャープ画像でしかポリープの形状をセグメンテーションすることしかできなかった []，しかしながら，臨床医が実際に使用する際にはノイズにロバストなセグメンテーションができることが重要である．この作業では，激しいモーションによって引き起こされた弱いエッジ領域やぼやけた領域からも内腔内の輪郭を正確に検出し，セグメンテーションをするはじめての研究である．

セマンティックセグメンテーションでは効果的に解決するために，低レベルな特徴マップを高レベルの特徴マップと組み合わせて正確なローカライズを可能にする設計が重要である [23,24]．したがって，U-net ライクなネットワークで，多くのセグメンテーションタスクが取り組まれてきた．この作業では，ノイズが多い内腔内の画像からポリープの形状を正確にセグメンテーションする効果的な方法を提案する．1) 画像中の複数の空間周波数特性に注目し，vanilla convolution を Octave Convolution(OctConv) に置き換え，Residual UNet に組み込むことで，階層的特徴をデコードする．2)U-net ライクなネットワークで一般的に使われているクロスエントロピー損失関数ではなく，Focal loss に変更することで，不均衡なセマンティッククラスを解消し，セグメンテーションの精度を向上させる．For example, 内視鏡のラベルでは，ポリープのピクセルは，non-polyp のピクセルよりも圧倒的に小さくクラス間の不均衡が生じている．

本論文の貢献は 2 つである．

(1) 空間周波数に着目した Convolution layer を使用したモデルのほうがノイズにロバストなセグメンテーションができる事を示した．

(2) 既存の手法では，非有益な画像として捨てられていたブラーフレームからもポリープの輪郭を正確に検出できる．

The remaining sections are organixzed as following: Section2 は，提案手法やセグメンテーションするためのネットワークモデル，損失関数，を説明する．section3 は，preprocessing や実験方法と実験結果を示す．section4 は，結論である．作成したコードは筆者のアカウントから確認できる．`https://github.com/mtakamat`.

https://polyp.grand-challenge.org/Databases/

2. PROPOSED METHOD

2.1. Octave Convolution

Octave Convolution (OctConv) layer [9] は，図のように，vanilla convolution layer を 2 つの解像度に分解して，特徴マップを異なる周波数マップ，低周波と高周波の特徴マップ $X = [X^H, X^L]$ に分割する．次に，output $Y = [Y^H, Y^L]$of the OctConv high and low frequency feature map は，それぞれ $Y^H = Y^{H\to H} + Y^{L\to H}$ and $Y^L = Y^{L\to L} + Y^{H\to L}$ と与えられる．$Y^{H\to H}, Y^{L\to L}$denote intra-frequency update, $Y^{H\to L}, Y^{L\to H}$denote inter-frequency communication.

$$\begin{aligned} Y^H &= f\left(X^H; W^{H\to H}\right) + upsample\left(f\left(X^L; W^{L\to H}\right), 2\right) \\ Y^L &= f\left(X^L; W^{L\to L}\right) + f\left(pool\left(X^H, 2\right); W^{H\to L}\right) \end{aligned}, \tag{1}$$

where $f(X; W)$ denotes a convolution with parameters W,$pool(X, k)$ is an average pooling operation with kernel size $k \times k$ and stride k. $upsample(X, k)$ is an up-sampling operation by a factor of k via nearest interpolation.In practice, the spatial dimensions of the X^L feature representation are divided by a factor of 2. This further helps each OctConv layer capture more contextual information from distant locations and can potentially improve recognition performance.

2.2. Residual Unit

The more layers of the network, the richer the features that can be extracted. However, blindly increasing the depth of the network will lead to gradient explosions or gradient dispersion. To overcome these problem, He et al. [] proposed the residual units. Each residual unit can be illustrated as a general form

$$\begin{aligned} \mathbf{y}_l &= h\left(\mathbf{x}_l\right) + \mathcal{F}\left(\mathbf{x}_l, \mathcal{W}_l\right) \\ \mathbf{x}_{l+1} &= f\left(\mathbf{y}_l\right) \end{aligned}, \tag{2}$$

where $\mathbf{x}_l$ and $\mathbf{x}_{l+1}$ are the input and output of the lth residual unit, $\mathcal{F}$ is the residual function, $f\left(\mathbf{y}_l\right)$ is the activation function, and h is an identity mapping function, that is $h\left(\mathbf{x}_l\right) = \mathbf{x}_l$. He et al. [] は，residual unit の様々な組み合わせの影響に関する詳細な実験を行い，図に示す full preactivation resiaual unit を提案した．本稿では，full preactivation resiaual unit を使用して Octave Residual Unet を構築する．

2.3. Octave Residual UNet

In this section, Octave UNet と residual neural network の両方を組み合わせた encoder-decoder based neural network architecture named Octave Residul UNet を提案する [, ,]. Residual UNet の大きな利点は，residual unit および Unet のスキップコネクションは，情報の伝播を促進し，はるかに少ないパラメータで neural network を設計できる上に，他のセメンティックセグメンテーションと同等の優れたパフォーマンスを達成できることである [road,skin]. In this paper, Octave Residual UNet is capable of extracting and decoding hierarchical multifrequency features for segmentation polyps from blur endoscope images. Octave Residual UNet consists of 9 levels architecture,the part of Encoder, Decoder (see Fig.2). 図のように複数のエンコーダブロックを積み重ねると，滑らかに変化する構造を表す低空間周波数成分と急速に変化する詳細を表す高空間周波数成分の両方を取得できる階層型マルチ周波数特徴を抽出できる．図に示すように，特徴をエンコーディングするプロセスで，特徴マップの空間次元が徐々に減少するため，the feature maps lose spatial details and location information. Skip connections is adopted to concatenate low level location information rich features to the inputs of decoder blocks. The stack of decoder block recovers spatial details and local information. All of the two parts are built with residual units that consist of two 3×3 OctConv layer and an identity mapping.

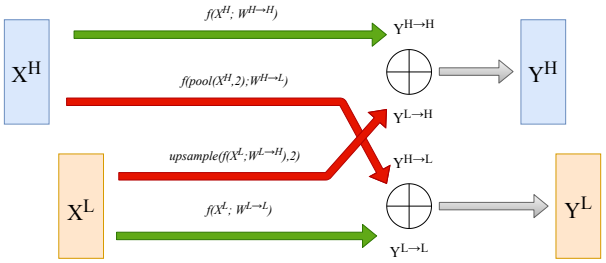

Fig. 1. OctConv;Green lines indicate intra-frequency,Red lines indicate inter-frequency

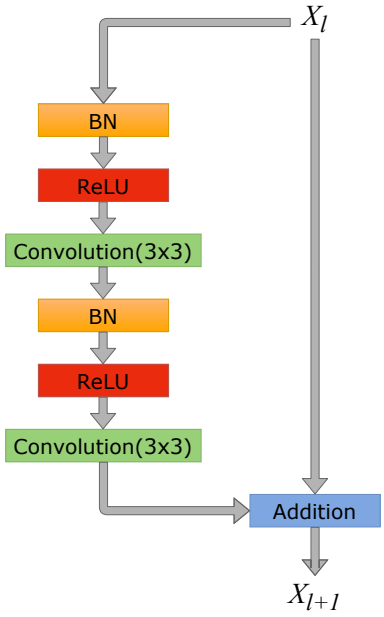

Fig. 2. Residual Unit

2.4. Loss Function

Focal loss は，学習時のクラス不均衡が one stage detector の検出精度に悪影響を与える問題を解決するためにクロスエントロピー損失を再構成したものである．クロスエントロピー損失でトレーニングされたセグメンテーションネットワークは，polyp pixel ではなく non-polyp pixel にむけてバイアスがかかるため，polyp pixel と non-polyp pixel のクラス不均衡が生じる．Focal loss can be illustrated as a general form

$$\mathrm{FL}\left(p_{\mathrm{t}}\right)=-\left(1-p_{\mathrm{t}}\right)^{\gamma} \log \left(p_{\mathrm{t}}\right), \tag{3}$$

where α_t is a balanced variant of the focal loss, γ is the smoothly adjusts the rate at which easy examples are down weighted.

3. EXPERIMENT AND RESULTS

3.1. Dataset and Pre-processing

1)*Dataset Description:*

- **CVC-ClinicDB** contains 612 images, where all images show at least one polyp. The segmentation labels obtained from 31 colorectal video sequences were acquired 23 patients.

- **CVC-ColonDB** contains 379 frames from 15 different colonoscopy sequences,where each sequence shows at least one polyp each.

- **ETIS-LaribPolypDB** contains 196 images, where all images show at one polyp.

2)*Preprocessing:*非有益な画像を多く含まれている画像のみからデータセットを作成するために，ラプラシアンフィルタのピクセル強度の分散から有益か非有益かを識別する．本稿では，OpenCV の *Variance_of_laplacian* 関数を使用する．分散の閾値は 100 である．
3)*Data Augmentation:* Data augmentation is used to artificially increase the size of the training dataset both image and label. Augmentation methods commonly employ transformations such as rotations, reflections, and elastic deformations, which produce training images that resemble one particular training example.

3.2. Implementation Details

All the models are implemented using Keras.Models are trained for 10 epochs using Adam optimizer,with a learning rate of 1e-4 and batch size 8.Experiments have been conducted with Nvidia Geforce1080 Ti GPU - 11GB RAM.

3.3. Evaluation Metrics

1)*Segmentation evaluation:* Jaccard index and Dice similarity score are the most commonly used evaluation metrics for segmentation.

$$Jaccard(A,B)=\frac{|A\cap B|}{|A\cup B|}, \tag{4}$$

$$\mathrm{Dice}(A,B)=\frac{2|A\cap B|}{|A|+|B|}, \tag{5}$$

2)*Intersection over Union(IoU):* Intersection over Union(IoU) is one of the standard measure performance measure for segmentation.Given a set of images, the IoU measure the similarity between the prediction and the ground truth per pixel.

$$IoU=\frac{TP}{TP+FN+FP}, \tag{6}$$

where TP, FP, FN denote the true positive,false positive and false negative numbers. This score is always between 0 and 1.

3.4. Results

4. CONCLUSION

5. REFERENCES