



Highlights

MingleNet: A Novel Dual Stacking Approach for Medical Image Segmentation

Hanson Gabriel Cendana,Dwayne Reinaldy,Bao Wei Chiam,Junnn Wei Chiang,Vincensius Hobart Wijaya,Kuan Y. Chang 

- MingleNet, using stacking ensemble learning, excels in medical image segmentation.
- MingleNet combines DoubleU-Net, DeepLabv3+, U-Net, and DeepLab for segmentation.
- MingleNet excels in polyp segmentation across benchmarks such as CVC-ColonDB.
- Dynamic image augmentation enhances model performance by introducing data variations.

MingleNet: A Novel Dual Stacking Approach for Medical Image Segmentation

Hanson Gabriel Cendana^a, Dwayne Reinaldy^a, Bao Wei Chiam^a, Jiunn Wei Chiang^a, Vincensius Hobart Wijaya^a and Kuan Y. Chang ^a

^aBiomedical AI Lab, Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, 20231, Taiwan (R.O.C.)

ARTICLE INFO

Keywords:

Image Segmentation
Stacking Ensemble Learning
Dynamic Image Augmentation
Polyp Segmentation
Deep Learning
Good Health & Well-being

ABSTRACT

Medical image segmentation holds significant importance in disease diagnosis and the formulation of treatment strategies. Ensemble learning, which combines multiple models or predictions, can improve accuracy and performance in medical image segmentation. MingleNet, our proposed model, employs a dual-layer ensemble learning approach, utilizing double-stacked models including DoubleU-Net, DeepLabv3+, U-Net, and DeepLab to generate masks. In MingleNet, the first layer's masks are averaged and concatenated with the original images to serve as input for the second layer. Additionally, we employ dynamic image augmentation—a novel, model-free single-image approach—to further enhance model performance. MingleNet undergoes evaluation on three polyp segmentation benchmarks: Kvasir-SEG, CVC-ClinicDB, and CVC-ColonDB. On Kvasir-SEG, MingleNet achieves 93.19 Dice, 87.24% IoU, 94.15% precision, 92.25% recall, and 97.87% accuracy. On CVC-ClinicDB, MingleNet achieves 95.99% Dice, 92.29% IoU, 96.08% precision, 95.90% recall, and 99.21% accuracy. On CVC-ColonDB, MingleNet achieves 94.33% Dice, 89.28% IoU, 95.89% precision, 92.83% recall, and 99.10% accuracy. MingleNet performs competitively across these benchmarks. Notably, it ranks 1st in Dice and IoU on the CVC-ColonDB benchmark, 2nd in Dice and IoU on CVC-ClinicDB, and 8th in Dice and 17th in IoU on the Kvasir-SEG benchmark.

1. Introduction

Medical image segmentation serves the critical purpose of enhancing the visibility of anatomical structures within images, which enables more precise analysis and diagnosis. By partitioning the image into different regions, it allows for extraction of relevant features and improves diagnostic accuracy. With the advancements in imaging technologies in recent years, like magnetic resonance imaging (MRI) scans and computed tomography (CT) scans, medical image segmentation has gained huge interest from researchers and practitioners [1, 2, 3].



The goal of medical image segmentation is to accurately describe the boundaries and contours of organs, tissues, or other structures of interest within medical images. The purpose of this process is to extract quantitative information, volumetric measurements, and spatial relationships, which are vital for clinical decision-making and precise intervention.

A wide range of segmentation techniques have been developed and employed in the field of medical imaging. Traditional methods include thresholding, region-based methods, clustering, and edge detection [2, 4, 5]. However, these methods frequently encounter challenges in dealing with the variability of medical images, often resulting in suboptimal outcomes.

In recent years, medical image segmentation has been revolutionized by the rise of deep learning-based techniques,

with convolutional neural networks (CNNs) playing a pivotal role in this transformation. CNNs have shown exceptional performance in various segmentation tasks by automatically learning hierarchical representations and capturing contextual information from medical images. Models such as U-Net [6], SegNet [7], DeepLab [8], TransUNet [9], UNet++ [10], FCN-Transformer [11], DUCK-Net [12] and Meta-Polyp [13] have been extensively used and adapted for medical image segmentation. U-Net [6], SegNet [7], DeepLab [8], UNet++ [10], DeepLabv3+ [14], and DoubleU-Net [15] are all deep CNN models for semantic segmentation, featuring an encoder (compression) path to capture context and a decoder (expansion) path for precise localization. From the encoder to the decoder, U-Net [6] transfers the entire feature maps, while SegNet [7] transfers only the pooling indices. U-Net uses skip connections to fuse low-level and high-level features, and up-convolutions to upsample the feature maps [6]. DeepLab uses atrous convolutions to enlarge the receptive field of the feature maps, and atrous spatial pyramid pooling to capture multi-scale information [8]. UNet++ is an improved version of U-Net, which uses nested and dense skip connections to enhance the feature fusion between the encoder and the decoder [10]. Instead, TransUNet is a transformer-based neural network that combines the strengths of both transformers and U-Net [9]. DeepLabv3+ is an advanced version of DeepLab featuring spatial attention mechanisms to enhance focus on relevant regions and improve localization [14]. DoubleU-Net is an extension of the original U-Net architectures, designed to further enhance segmentation performance by incorporating dual encoding and decoding pathways to capture more intricate features [15]. FCN-Transformer combines the strengths of fully convolutional networks (FCNs) and transformers

*Corresponding author

 kchang@ntou.edu.tw (K.Y.C. )

to capture fine-grained details and long-range relationships within the image [11]. DUCK-Net, a CNN architecture tailored for accurate polyp segmentation, achieves excellent performance even with limited training data by using custom DUCK Block and residual downsampling [12]. Meta-Polyp is also a model designed for poly segmentation. It combines the Meta-Former architecture with UNet and utilizes a new Convformer block to capture both fine details and global context for accurate yet efficient polyp detection [13].

Ensemble learning is a machine learning technique that integrates the predictions of multiple models to improve a single model by reducing variance and exploiting the advantages of each model [16]. Stacking ensemble learning, a widely used technique, involves training multiple models (usually of different types) on the same data. Subsequently, a meta-learner—a separate model—is employed to learn the optimal way to combine the predictions from these base models [17].

In this study, we explored whether integrating ensemble learning techniques could substantially improve the performance of medical image segmentation models, especially in polyp segmentation. To choose our base model, we carefully reviewed the existing literature, focusing on results such as Kvasir-SEG [18], CVC-ClinicDB [19], and CVC-ColonDB [20] benchmarks. We selected DeepLabv3+ [14] and DoubleU-Net [15] specifically because both have high performance in segmentation tasks and both models are the best models derived from two different models, DeepLab [8] and U-Net [6]. We then leveraged the uniqueness of ensemble learning to enhance performance. By combining diverse models, we captured complementary patterns and robustness across various datasets, capitalizing on each model's unique architecture and strengths.

2. Related Works

2.1. DivergentNets

DivergentNets is a medical image segmentation technique that uses an ensemble of multiple high-performing image segmentation architectures [21]. The model combines the TriUNet segmentation model with an ensemble of well-known segmentation models, namely UNet++ [10], FPN [22], DeepLabv3 [23], and DeepLabv3+ [14]. The TriUNet model takes a single image as input and then passes it into two distinct U-Net models [21]. The outputs of these two models are combined and then processed through a third U-Net model to generate the final segmentation masks [21].

In the EndoCV2021 challenge, the TriUNet architecture used in DivergentNets was the winning model in terms of segmentation accuracy and generalization [21]. In contrast to the original TriUNet method, DivergentNets employs an ensemble of five intermediate models (TriUNet [21], UNet++ [10], FPN [22], DeepLabv3 [23], and DeepLabv3+ [14]), which are trained separately and then combined by averaging the pixels between each mask [21].

In comparison, DivergentNets employs five different medical image segmentation models and employ single layer

of models, while MingleNet uses four models, DeepLabv3+ [14], DoubleU-Net [15], DeepLab [8], and U-Net [6]. Furthermore, MingleNet employs a stacking approach with multiple layers of models, utilizing two layers of deep learning models instead of the single layer used in DivergentNets. We employ the identical method, specifically the averaging technique to combine predictions.

3. Material and Methods

3.1. Polyp Image Datasets

3.1.1. Kvasir-SEG

The Kvasir-SEG v2 dataset [18] which consists of 1000 gastrointestinal segmented polyp images and corresponding segmentation masks, was manually annotated and verified by medical experts. The image resolution ranges from 332 x 487 to 1920 x 1072 pixels.

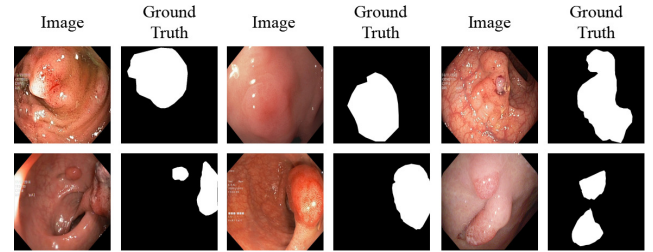


Figure 1: Example of Images and Masks from Kvasir-SEG. Each of the six images has a mask on its right side.

3.1.2. CVC-ClinicDB

CVC-ClinicDB [19] is an open-access dataset of 612 gastrointestinal segmented polyp images and their ground truth from 31 colonoscopy videos at a resolution of 384 x 288 pixels.

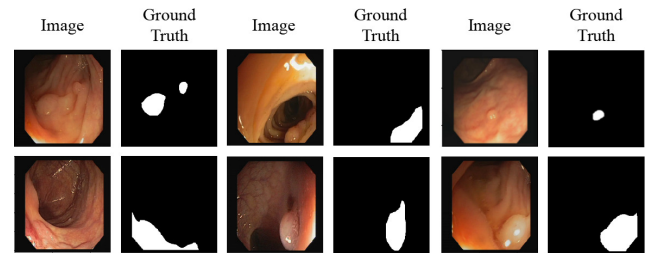


Figure 2: Example of Images and Masks from CVC-ClinicDB. Each of the six images has a mask on its right side.

3.1.3. CVC-ColonDB

CVC-ColonDB [20] is an open-access dataset of 380 colon-segmented polyp images and masks from 13 patients' colonoscopy videos at 500 x 574 pixels.

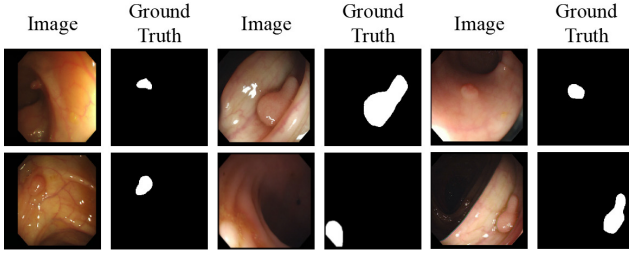


Figure 3: Example of Images and Masks from CVC-ColonDB. Each of the six images has a mask on its right side.

In this study, we rescaled Kvasir-SEG and CVC-ColonDB images to 352 x 352 pixels and CVC-ClinicDB images to 256 x 256 pixels.

3.2. Dynamic Image Augmentation

We proposed a novel model-free single-image approach called Dynamic Image Augmentation (DIA), which adapts and modifies the method from DUCK-Net [12], itself based on the work of Sanderson and Matuszewski [11], to generate new augmented images before each training epoch. We implemented this method using the Albumentations library [24]. Specifically, we modified the DUCK-Net method by introducing randomized horizontal and vertical flips, as well as Gaussian blur, each with a 0.5 probability.

1. Horizontal flips and vertical flips with a probability of 0.5,
2. Color jitter with a brightness range of [0.6, 1.6], a fixed contrast of 0.2, a saturation factor of 0.1, and a hue factor of 0.01,
3. Affine transformation with rotations within the range of $[-180^\circ, 180^\circ]$, horizontal and vertical translations within $[-0.125, 0.125]$, scaling with a magnitude within [0.5, 1.5], and shearing with an angle within $[-22.5^\circ, 22.5^\circ]$.
4. Gaussian blur with blur limit of [25, 25], sigma limit of [0.001, 2.0], and probability of 0.5.

Techniques such as flipping, adjusting brightness, rotation [2, 25], and blurring introduce diversity into the dataset [2], as shown in Fig. 4. We use probability-based horizontal and vertical flips to enhance data variation in each epoch, thereby improving model generalization. This means that images within an epoch can exhibit only horizontal flips, only vertical flips, both horizontal and vertical flips, or no flips at all. For color jitter and affine transformations, we define specific value ranges for image augmentation. Consequently, each epoch experiences different color jitter and affine transformations, as these parameters are randomly selected from their respective ranges. Lastly, Gaussian blur is applied probabilistically, allowing images in each epoch to exhibit either blur or no blur.

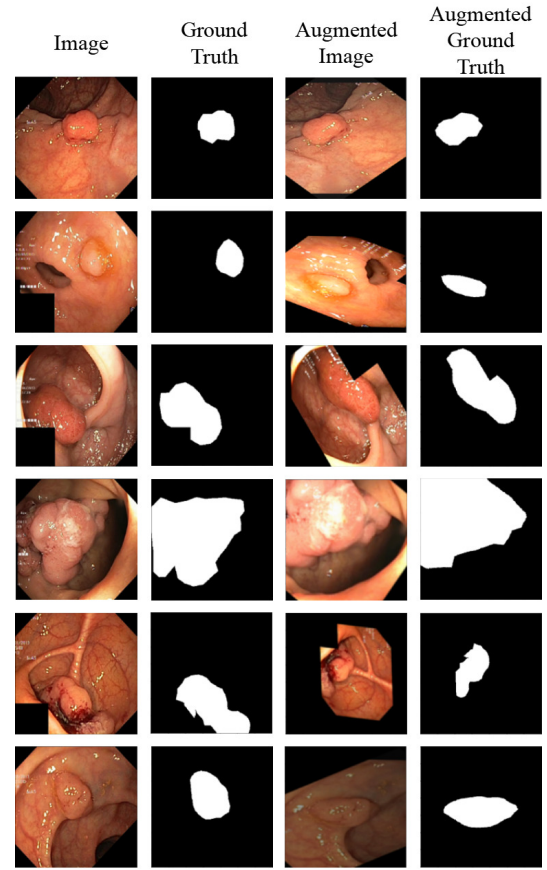


Figure 4: Example of images, masks, augmented images, and augmented masks. Each of the six images is followed by a mask, an augmented image, and its corresponding augmented mask.

3.3. Deep Learning Models for Image Segmentation

3.3.1. DeepLabv3+

DeepLabv3+ is an enhanced version of DeepLab, a semantic segmentation model. DeepLab utilizes dilated convolutions to process images, with outputs fine-tuned by a fully connected conditional random field (CRF) and bilinear interpolation. DeepLabv3+ incorporates semantic information from its encoder and recovers precise object boundaries through the decoder. The encoder uses atrous convolution to extract features at any resolution, while the decoder refines object boundaries. DeepLabv3+ also improves speed and accuracy by adapting the ResNet101 model [14].

3.3.2. DoubleU-Net

DoubleU-Net is a semantic segmentation model that enhances the process by using two U-Net architectures. It contains five components: two U-Net networks, VGG19 [26], a squeeze-and-excite block, and atrous spatial pyramid pooling (ASPP). The first U-Net utilizes VGG19 [26] as an encoder and ASPP as the input of the decoder. The squeeze-and-excite block improves the feature maps in the encoder of the first U-Net and those of both decoders. The output of the first U-Net and the corresponding image's mask are

multiplied by the input pictures. The second U-Net also utilizes ASPP and the squeeze-and-excite block. Finally, the result is obtained through concatenation of the outputs from both U-Nets [15].

3.3.3. Comparison between DeepLabv3+ and DoubleU-Net

DeepLabv3+ [14] uses atrous convolutions to capture multiscale contextual information, excelling in segmentation tasks where context is crucial. DoubleU-Net [15] enhances the U-Net framework with a dual pathway for feature fusion across multiple scales, improving its ability to handle various object sizes. Both models use ASPP and incorporate backbones in their architecture: DoubleU-Net [15] uses VGG19, while DeepLabv3+ [14] uses ResNet101.

3.4. Stacking Ensemble Learning

Stacking ensemble learning is a method that combines predictions from multiple models into a meta-model for better results, as shown in Fig. 5. Ensemble learning, in general, is a powerful machine learning technique that merges various models to outperform a single model [17]. The core idea is to use the diversity of these models to reduce the limitations and biases of individual models, thereby improving overall performance and generalization [17].

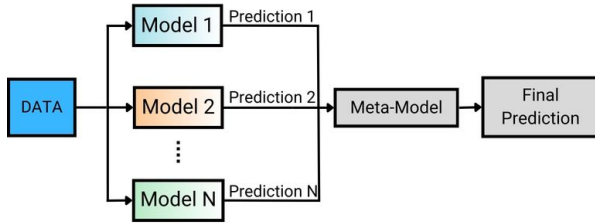


Figure 5: Concept of Stacking Ensemble Learning

3.5. MingleNet Architecture

MingleNet is a semantic segmentation model that uses stacking ensemble learning to combine multiple model layers, as illustrated in Fig. 6. The first layer of MingleNet consists of two different deep learning models, namely DeepLabv3+ [14] and DoubleU-Net [15], which take the original image with three channels (RGB) as input. Next, a mask is generated for each channel along with an average mask across all three channels. The output from the first layer, therefore, consists of four channels: the individual RGB channels and the average mask.

The second layer consists of another two deep learning models namely U-Net [6] and DeepLab [8], which take the output of the first layer as an input. We selected U-Net [6] and DeepLab [8] for the second layer instead of DeepLabv3+ [14] and DoubleU-Net [15] because DeepLabv3+ [14] employs a ResNet101 backbone, and DoubleU-Net [15] utilizes a VGG19 backbone, both of which lack support for a 4-channel input. The final output is the average mask computed across the four channels from the second layer.

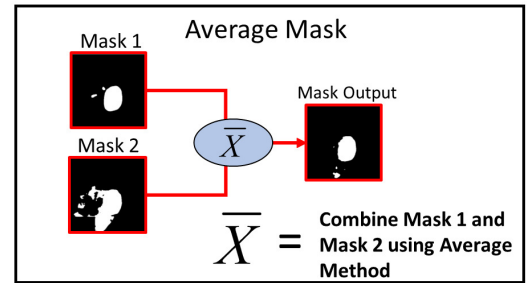
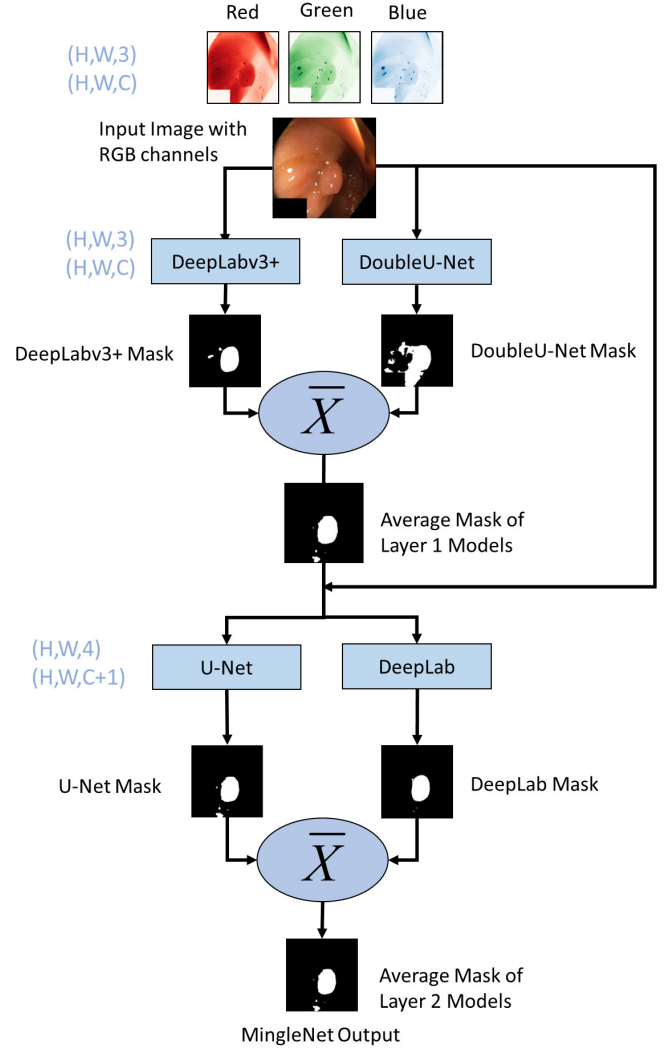


Figure 6: MingleNet architecture. Diagram of the proposed ensemble learning technique

Each model in the first and second layer was trained with a batch size of 4 for 600 and 400 epochs, respectively. Both layers utilized the AdamW [27] optimizer and binary cross entropy loss with a weight decay of 0.004. However, the learning rates differed. The first layer used a learning rate of 1×10^{-4} , while the second layer used that of 1×10^{-5} . This adjustment aims to minimize the likelihood of the models bypassing optimal parameters, facilitating a more effective convergence to the minimum of the loss function. Additionally, DIA was applied in every epoch for the first

layer, and the best-performing model was saved at the end. However, for the second layer, DIA was not performed due to the time complexity involved in its implementation.

3.6. Implementation Details

We experimented on an Intel Core i7-12700 processor with an RTX 3080 TI GPU and 32GB of RAM at the Biomedical AI Lab at National Taiwan Ocean University to run Tensorflow 2.14.0 and Python 3.10.13.

3.7. Evaluation

For evaluation, we use the following performance metrics: Dice Coefficient, Intersection over Union (IoU), Precision, Recall, and Accuracy. The formula are presented below:

$$\text{DiceCoefficient} = \frac{2TP}{2TP + FP + FN} \quad (1)$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

All the metrics used in this study are based on binary classification, where TP, FP, TN, and FN correspond to true positive, false positive, true negative, and false negative, respectively [28]. These metrics ensure that our evaluation directly compares predicted results to the ground truth represented by the true mask.

4. Results

4.1. Comparison of image augmentation across datasets

Table 1 demonstrates that image augmentation enhances all semantic segmentation models substantially. On Kvasir-SEG dataset, DIA consistently outperforms the Duck-Net augmentation method across all deep learning models. For example, with DoubleU-Net [15], DIA achieves a Dice score of 92.58% and an IoU of 86.18%, compared to 90.68% Dice and 82.95% IoU with the DUCK-Net augmentation.

However, the DUCK-Net Augmentation surpasses DIA in several segmentation models on the CVC-ClinicDB and CVC-ColonDB datasets. For example, on CVC-ClinicDB, the DUCK-Net augmentation achieves a Dice score of 97.14% and an IoU of 94.44% with DeepLabv3+ [16], compared to DIA's 95.19% Dice and 90.83% IoU. Similarly, on the CVC-ColonDB dataset, the DUCK-Net Augmentation achieves a Dice score of 90.81% and an IoU of 83.16% with DeepLabv3+ [16], compared to DIA's 90.37% Dice and 82.44% IoU.

4.2. MingleNet vs other models on Kvasir-SEG

Table 2 summarizes the performance metrics of several semantic segmentation models on the Kvasir-SEG benchmark. At the first layer of MingleNet, DoubleU-Net surpasses DeepLabv3+ in Dice, IoU, and Recall metrics, while U-Net outperforms DeepLab at the subsequent layer after averaging the predicted masks from DoubleU-Net and DeepLabv3+. MingleNet's final output enhances Dice by 0.61% and IoU by 1.13% compared to DoubleU-Net at the first layer. Meanwhile, Meta-Polyp [13] leads with Dice at 95.90% and IoU at 87.24%, whereas DUCK-Net tops [12] in Precision (96.28%), Recall (93.79%), and Accuracy (98.42%). Despite not exceeding DUCK-Net and Meta-Polyp, MingleNet demonstrates strong performance, confirming its effectiveness in polyp segmentation tasks.

Fig. 7 illustrates the progressive results of MingleNet on the Kvasir-SEG benchmark. MingleNet combines base models (DoubleU-Net and DeepLabv3+), averaging their predictions and passing them to a second layer (U-Net and DeepLab). The resulting final output effectively identifies polyp shapes with fewer false positives than the base models. However, in the first of the three cases, DeepLab at the second layer produces better masks with fewer false positives compared to MingleNet.

4.3. MingleNet vs other models on CVC-ClinicDB

Table 3 summarizes semantic segmentation model performance on the CVC-ClinicDB benchmark. At the first layer of MingleNet, DoubleU-Net surpasses DeepLabv3+ in Dice and IoU by 0.70% and 1.28% respectively, while U-Net outperforms DeepLab at the second layer after averaging the predictions from DoubleU-Net and DeepLabv3+. MingleNet achieves the highest Dice (95.99%), IoU (92.29%), Recall (95.95%), and Accuracy (99.21%) despite not surpassing FCN-Transformer [11] in Precision (96.59%). It validates MingleNet's efficacy in semantic segmentation.

Fig. 8 illustrates the progressive results of MingleNet on the CVC-ClinicDB benchmark. Starting with base models (DoubleU-Net and DeepLabv3+), predictions are averaged and passed to a second layer (U-Net and DeepLab). MingleNet's final output identifies polyp shapes, successfully reducing false positives compared to the base models in all three cases.

4.4. MingleNet vs other models on CVC-ColonDB

Table 4 summarizes semantic segmentation model performance on the CVC-ColonDB benchmark. The first layer of MingleNet uses DoubleU-Net and DeepLabv3+ to surpass DeepLabv3+ alone in Dice, IoU, and Recall. The second layer of MingleNet averages predictions as an input to make DeepLab outperform U-Net. MingleNet's final output enhances Dice by 0.63% and IoU by 1.21% compared to DoubleU-Net. It achieves highest scores in Dice (94.34%), IoU (89.28%), and Precision (95.89%), while DUCK-Net [12] leads in Recall (93.92%) and Accuracy (99.29%). MingleNet demonstrates efficacy in semantic segmentation, excelling in Dice, IoU, and Precision.

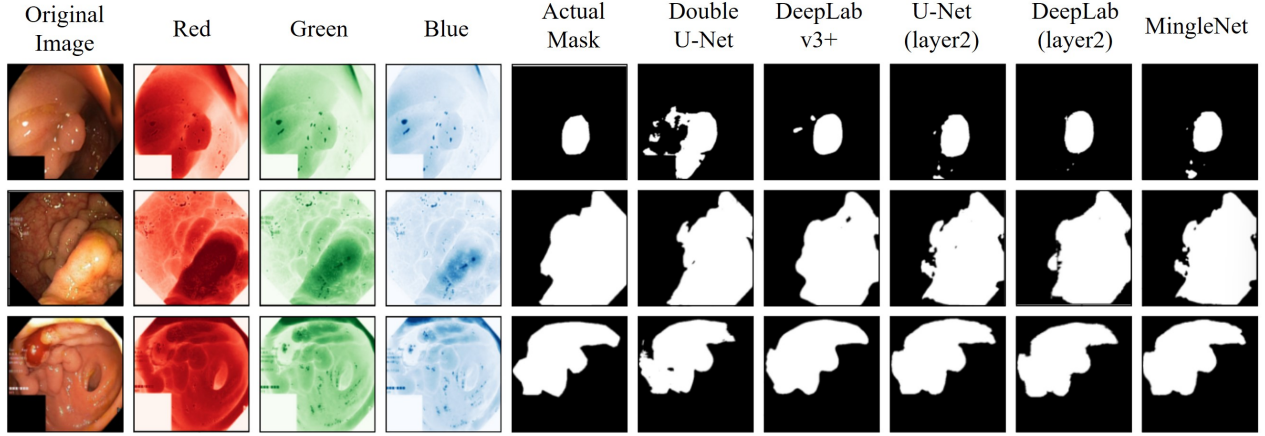


Figure 7: Segmented polyp masks of MingleNet components on Kvasir-SEG benchmark. The figure displays the actual masks, RGB channels, and the results of each segmentation from the first layer (DoubleU-Net and DeepLabv3+), the second layer (U-Net, DeepLab), and the final output of MingleNet.

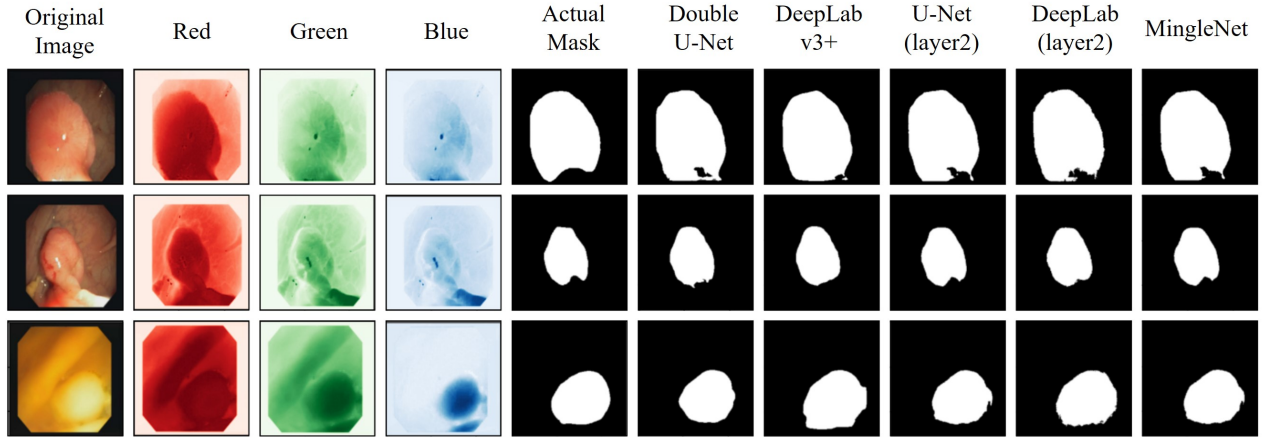


Figure 8: Segmented polyp masks of MingleNet components on the CVC-ClinicDB benchmark. The figure displays the actual mask, RGB channel, and the result of each segmentation from the first layer (DoubleU-Net and DeepLabv3+), the second layer (U-Net, DeepLab), and the final output.

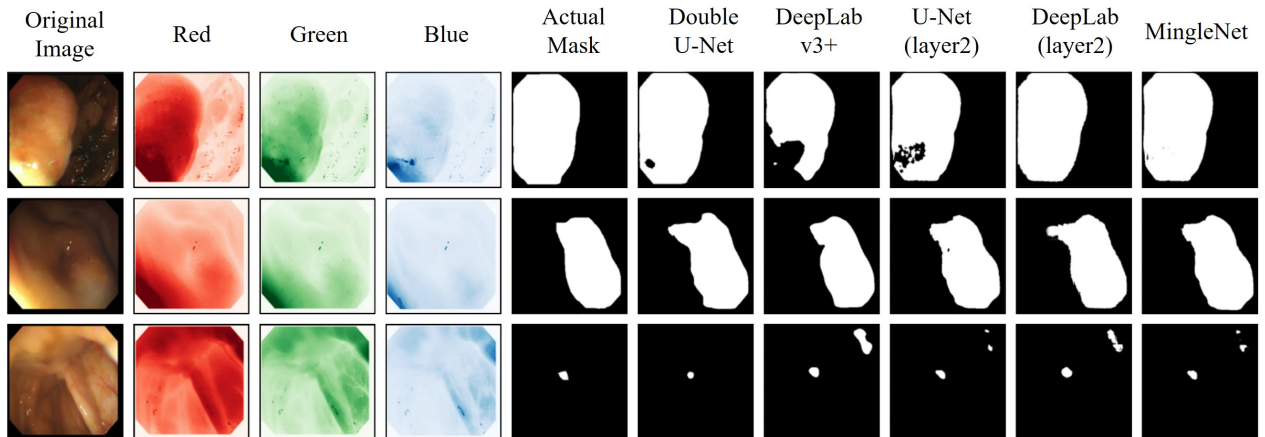


Figure 9: Segmented polyp masks of MingleNet components on CVC-ColonDB benchmark. The figure displays the actual mask, RGB channel, and the result of each segmentation from the first layer (DoubleU-Net and DeepLabv3+), the second layer (U-Net, DeepLab), and the final output.

Table 1

Image augmentation methods on model performance across datasets

Dataset	Model	Without augmentation		DUCK-Net Augmentation		DIA	
		Dice	IoU	Dice	IoU	Dice	IoU
Kvasir-SEG	U-Net [6]	83.59%	74.28%	80.93%	67.98%	86.73%	76.57%
	DeepLab [8]	75.41%	66.21%	70.91%	54.93%	87.61%	77.95%
	DeepLabv3+ [14]	87.20%	77.30%	91.62%	84.54%	91.86%	84.94%
	DoubleU-Net [15]	83.62%	71.86%	90.68%	82.95%	92.58%	86.18%
CVC-ClinicDB	U-Net [6]	91.06%	81.51%	92.17%	85.48%	94.25%	89.13%
	DeepLab [8]	86.72%	74.09%	91.24%	83.89%	93.35%	87.53%
	DeepLabv3+ [14]	93.76%	88.25%	97.14%	94.44%	95.19%	90.83%
	DoubleU-Net [15]	93.66%	88.08%	97.10%	94.36%	95.89%	92.11%
CVC-ColonDB	U-Net [6]	86.50%	76.20%	90.62%	82.86%	90.46%	82.15%
	DeepLab [8]	75.41%	66.21%	91.04%	83.56%	90.87%	83.28%
	DeepLabv3+ [14]	89.60%	81.16%	90.81%	83.16%	90.37%	82.44%
	DoubleU-Net [15]	90.42%	82.51%	92.90%	86.75%	93.74%	88.21%

Table 2

Comparison of model performance on the Kvasir-SEG dataset

Model	Dice	IoU	Precision	Recall	Accuracy
U-Net [6]	83.59%	74.28%	86.45%	76.14%	93.94%
FCN-Transformer(pre-trained) [11]	92.20%	85.54%	92.38%	92.03%	97.49%
DUCK-Net(pre-trained) [12]	95.02%	90.51%	96.28%	93.79%	98.42%
Meta-Polyp(pre-trained) [13]	95.90%	92.10%	-	-	-
TransResU-Net(pre-trained) [29]	88.84%	82.14%	91.06%	90.22%	96.51%
MingleNet					
Layer 1:DoubleU-Net	92.58%	86.18%	92.28%	92.87%	97.64%
Layer 1:DeepLabv3+	91.86%	84.94%	93.76%	90.03%	97.48%
Layer 2:U-Net	93.13%	87.14%	93.92%	92.35%	97.84%
Layer 2:Deeplab	92.69%	86.37%	95.12%	90.37%	97.74%
Final: MingleNet	93.19%	87.24%	94.15%	92.25%	97.87%

Fig. 9 shows the progressive result in MingleNet on CVC-ColonDB test set. Starting from base models consisting of DoubleU-Net and DeepLabv3+, whose predictions are averaged and fed pass into the second layer comprising U-Net and DeepLab. Then MingleNet Output results from the averaged prediction masks of U-Net and DeepLab. Fig. 9 reveals that the resulting output shows that MingleNet was

able to identify the shape of the polyp objects, although there are some issues in Row 3 where it successfully decreased its false negatives in the polyp object area but still showed false positives in the outer parts.

Table 3

Comparison of model performance on the CVC-ClinicDB dataset

Model	Dice	IoU	Precision	Recall	Accuracy
U-Net [6]	91.06%	81.51%	95.26%	87.22%	97.77%
FCN-Transformer(pre-trained) [11]	88.00%	78.58%	96.59%	80.82%	96.45%
DUCK-Net(pre-trained) [12]	94.78%	90.09%	94.68%	94.89%	99.07%
MingleNet					
Layer 1:DoubleU-Net	95.89%	92.11%	95.61%	96.17%	99.18%
Layer 1:DeepLabv3+	95.19%	90.83%	95.72%	94.67%	99.05%
Layer 2:U-Net	95.94%	92.20%	96.17%	95.71%	99.20%
Layer 2:Deeplab	95.56%	91.51%	94.69%	96.46%	99.11%
Final Output	95.99%	92.29%	96.08%	95.90%	99.21%

Table 4

Comparison of model performance on the CVC-ColonDB dataset

Model	Dice	IoU	Precision	Recall	Accuracy
U-Net [6]	86.50%	76.2%	94.34%	79.94%	97.99%
FCN-Transformer(pre-trained) [11]	90.73%	83.04%	91.07%	90.40%	98.99%
DUCK-Net (pre-trained) [12]	93.53%	87.85%	93.14%	93.92%	99.29%
Meta-Polyp(pre-trained) [13]	86.70%	79.00%	-	-	-
MingleNet					
Layer 1:DoubleU-Net	93.74%	88.21%	94.94%	92.56%	99.00%
Layer 1:DeepLabv3+	90.37%	82.44%	96.29%	85.14%	98.53%
Layer 2:U-Net	93.60%	87.97%	95.88%	91.42%	98.99%
Layer 2:Deeplab	93.91%	88.52%	94.74%	93.08%	99.02%
Final Output	94.33%	89.28%	95.89%	92.83%	99.10%

5. Discussion

In this study, we introduce MingleNet, a deep learning model that leverages ensemble learning for semantic segmentation tasks. MingleNet consistently outperformed individual base models. Ensemble learning capitalizes on the collaborative strength of these base models, each contributing unique insights. While individual base models excel in certain aspects, they struggle with the full complexity of benchmarks. Ensemble learning bridges these gaps by leveraging diverse patterns captured by each model, resulting in a better understanding of the data.

In fact, MingleNet excels in polyp segmentation. According to paperswithcode [30], MingleNet ranks 8th in Dice (93.19%) and 17th in IoU (87.24%) for the Kvasir-SEG benchmark. In the CVC-ClinicDB benchmark, it achieves 2nd in Dice (95.99%) and IoU (92.29%), while in CVC-ColonDB, it leads with 1st in Dice (94.34%) and IoU (89.28%).

We find that incorporating weaker models such as U-Net and DeepLab at the second layer of MingleNet improves the results of the first layer slightly. The second layer benefits from including segmentation averaged masks generated by the first layer. These averaged masks, added as a fourth channel to the dataset, provide additional context that enhances the performance of the second layer models.

Our approach parallels DivergentNets [21], which combines an ensemble of well-known segmentation models (including UNet++ [10], FPN [22], DeepLabv3 [23], and DeepLabv3+ [14]) with TriUNet [21] to tackle generalization in polyp segmentation. While both MingleNet and DivergentNets [21] utilize ensemble learning to improve segmentation accuracy, MingleNet distinguishes itself by employing a two-layer stacking approach. In contrast, DivergentNets focuses on combining the strengths of multiple well-known segmentation models in a single ensemble to produce more generalizable segmentation masks. Both methods illustrate the power of ensemble learning, but MingleNet's two-stage stacking process adds an extra dimension of refinement to the training process.

In addition, we demonstrate that DIA enhances the performance of deep learning semantic segmentation models.

DIA improves model performance by incorporating augmented images before each training epoch, even when dealing with larger datasets like Kvasir-SEG (with 1000 samples) that may include blurred images due to augmentation. Despite the Gaussian blur, the diversity within Kvasir-SEG ensures essential features are preserved. Conversely, smaller datasets like CVC-ColonDB (with only 480 samples) suffer more from the Gaussian blur effects, emphasizing the need to preserve fine details. DIA, including Gaussian blur, benefits Kvasir-SEG due to its larger sample size. In contrast, DUCK-Net Augmentation (without Gaussian blur) [12] is more effective for datasets like CVC-ColonDB, maintaining optimal performance. DIA provides sufficient variability without compromising the integrity of important image features, ensuring optimal performance in such contexts.

Furthermore, the success of MingleNet on benchmarks such as Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB aligns with the performance of DivergentNets in the EndoCV 2021 segmentation challenge, reinforcing the notion that ensemble approaches, when carefully constructed, can significantly advance the state of the art in medical image analysis. The complementary nature of these models suggest that combining ensemble learning strategies, as seen in both MingleNet and DivergentNets, offers a robust pathway for enhancing segmentation accuracy and generalization across various datasets and challenges.

6. Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. Acknowledgments

The work was supported by the Summer Internship Program for College Students of R.O.C. under NTOU-112-E-108. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- [1] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, Y. Qiu, Recent advances and clinical applications of deep learning in medical image analysis, *Medical Image Analysis* 79 (2022) 102444. doi:10.1016/j.media.2022.102444.
- [2] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, A. K. Nandi, Medical image segmentation using deep learning: A survey, *IET Image Processing* 16 (5) (2022) 1243–1267. doi:10.1049/ipr2.12419.
- [3] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, Y. Xie, From cnn to transformer: A review of medical image segmentation models, *Imaging Informatics in Medicine* 37 (4) (2024) 1529–1547. doi:10.1007/s10278-024-00981-7.
- [4] R. M. Haralick, L. G. Shapiro, Image segmentation techniques, *Computer Vision, Graphics, and Image Processing* 29 (1) (1985) 100–132. doi:10.1016/S0734-189X(85)90153-7.
- [5] D. L. Pham, C. Xu, J. L. Prince, Current methods in medical image segmentation, *Annual Review of Biomedical Engineering* 2 (1) (2000) 315–337, PMID: 11701515. doi:10.1146/annurev.bioeng.2.1.315.
- [6] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.
- [7] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (12) (2017) 2481–2495. doi:10.1109/TPAMI.2016.2644615.
- [8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (4) (2018) 834–848. doi:10.1109/TPAMI.2017.2699184.
- [9] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, X. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, Y. Zhou, Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers, *Medical Image Analysis* 97 (2024) 103280. doi:10.1016/j.media.2024.103280.
- [10] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, A. Madabhushi (Eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer International Publishing, Cham, 2018, pp. 3–11. doi:10.1007/978-3-030-00889-5_1.
- [11] E. Sanderson, B. J. Matuszewski, Fcn-transformer feature fusion for polyp segmentation, in: G. Yang, A. Aviles-Rivero, M. Roberts, C.-B. Schönlieb (Eds.), *Medical Image Understanding and Analysis*, Springer International Publishing, Cham, 2022, pp. 892–907. doi:10.1007/978-3-031-12053-4_65.
- [12] R.-G. Dumitru, D. Peteleaza, C. Craciun, Using duck-net for polyp image segmentation, *Scientific Reports* 13 (1) (2023) 9803. doi:10.1038/s41598-023-36940-5.
- [13] Q. Trinh, Meta-polyp: A baseline for efficient polyp segmentation, in: *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 742–747. doi:10.1109/CBMS58004.2023.00312.
- [14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), *Computer Vision – ECCV 2018*, Springer International Publishing, Cham, 2018, pp. 833–851. doi:10.1007/978-3-030-01234-2_49.
- [15] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, 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)*, IEEE, 2020, pp. 558–564. doi:10.1109/CBMS49503.2020.00111.
- [16] D. Opitz, R. Maclin, Popular ensemble methods: an empirical study, *J. Artif. Int. Res.* 11 (1) (1999) 169–198. doi:10.1613/jair.614.
- [17] P. Proskura, A. Zaytsev, Effective training-time stacking for ensembling of deep neural networks, in: *Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition, AIPR '22*, Association for Computing Machinery, New York, NY, USA, 2023, p. 78–82. doi:10.1145/3573942.3573954.
- [18] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, W. De Neve (Eds.), *MultiMedia Modeling*, Springer International Publishing, Cham, 2020, pp. 451–462. doi:10.1007/978-3-030-37734-2_37.
- [19] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, *Computerized Medical Imaging and Graphics* 43 (2015) 99–111. doi:10.1016/j.compmedimag.2015.02.007.
- [20] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdal, A. Courville, A benchmark for endoluminal scene segmentation of colonoscopy images, *Journal of Healthcare Engineering* 2017 (2017) 4037190. doi:10.1155/2017/4037190.
- [21] V. L. Thambawita, S. Hicks, P. Halvorsen, M. Riegler, Divergent-nets: medical image segmentation by network ensemble, in: *Proceedings of the CEUR Workshop on Computer Vision in Endoscopy 2021*, Vol. 2886 of CEUR Workshop Proceedings, 2021, pp. 27–38, <https://ceur-ws.org/Vol-2886/paper3.pdf>.
- [22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection (2017). doi:10.1109/CVPR.2017.106.
- [23] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation (2017). doi:10.48550/arXiv.1706.05587.
- [24] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A. A. Kalinin, Albumentations: Fast and flexible image augmentations, *Information* 11 (2) (2020). doi:10.3390/info11020125.
- [25] M. Xu, S. Yoon, A. Fuentes, D. S. Park, A comprehensive survey of image augmentation techniques for deep learning, *Pattern Recognition* 137 (2023) 109347. doi:10.1016/j.patcog.2023.109347.
- [26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: *3rd International Conference on Learning Representations (ICLR 2015)*, Computational and Biological Learning Society, 2015, pp. 1–14, <https://ora.ox.ac.uk/objects/uuid:60713f18-a6d1-4d97-8f45-b60ad8aebbbe>.
- [27] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: *7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019, <https://dblp.org/rec/conf/iclr/LoshchilovH19.html>.
- [28] J. Lever, M. Krzywinski, N. Altman, Classification evaluation, *Nature Methods* 13 (8) (2016) 603–604. doi:10.1038/nmeth.3945.
- [29] N. K. Tomar, A. Shergill, B. Rieders, U. Bagci, D. Jha, Transresu-net: A transformer based resu-net for real-time colon polyp segmentation, in: *2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2023, pp. 1–4. doi:10.1109/EMBC40787.2023.10340572.
- [30] Meta Platforms, Ins., *Papers with Code: Medical Image Segmentation*, accessed: 28 August 2024, <https://paperswithcode.com/task/medical-image-segmentation> (n.d.).