Abstract

Semantic image segmentation is the process of labeling each pixel of an image with its corresponding class. U-Net is a popular Convolutional Neural Network architecture, originally proposed for biomedical image segmentation, but now widely used in a range of additional segmentation tasks. Many architectures have been built on top of the UNet framework, such as U-Net++ [12],unet^2[11], *DoubleUnet*[7], *MultiResUNet*[6], *and many others.*

In this paper, we propose a new architecture called MGUnet, which is a multigrid CNN consisting of n concatenated interconnected UNet components. MGUnet can be used for classification and weakly supervised segmentation tasks as well. We evaluate MGUnet on medical segmentation datasets covering various medical techniques such as dermatoscopy, colonoscopy, and microscopy, and compare it with the UNet and DoubleUnet architectures. Our results show that MGUnet outperforms the UNet model on all datasets except for one. In addition we compared the model with the DoubleUnet architecture which has been shown to outperform several other improved UNet architectures, such as Multi-ResUNet, Unet++ in terms of dice coefficient and IOU loss.

1 Introduction

Image segmentation has a long history and has traditionally been performed manually by experts with prior knowledge and experience in the field. The manual approach involved the use of techniques such as thresholding, edge detection, and region growing to identify and isolate different regions or objects within an image. With advancements in computer vision and machine learning, automated approaches for image segmentation have become more prevalent and are now used widely in various applications. These automated approaches rely on sophisticated algorithms [3],[8], [**?**] to perform the segmentation, and are typically more efficient and less prone to human error compared to the manual approach. In recent years, the common approach for solving image segmentation problems has shifted towards treating the task as a pixel-wise classification problem. This is accomplished by obtaining the classification information of each pixel through the use of Convolutional Neural Networks (CNNs [16]). This approach has become prevalent due to the improved accuracy and efficiency achieved through the use of deep learning techniques.

2 Related work

Unet [17]which was specifically designed for medical image segmentation tasks, and thanks to its popularity in this field, it is now used in a range of additional segmentation, and detection tasks. The architecture was designed to handle the challenges of medical image segmentation, such as small objects, complex structures, small training datasets, and class imbalance. The architecture is a convolutional neural network, consists of two main components: encoder and decoder. The encoder is responsible for reducing the spatial dimension of the input. This lower-resolution representation retains important information about the local region surrounding each pixel, consequently the amount of channels of the original increases during this phase. Then the last decoder layer passes its output to the bottleneck layer which helps to reduce the number of channels in the feature map and compresses the high-level features into a smaller, more manageable representation. Then the decoder pathaway takes the output from the bottleneck layer and gradually increases the spatial dimensions of the feature map through a series of upsampling, concatenation - which is one of , and convolutional operations.

The contracting path is typically composed of several blocks of convolutional layers, each followed by a max pooling layer to reduce the spatial dimensions. The expansive path, on the other hand, is composed of several blocks of convolutional layers, each followed by a transposed convolutional layer (also known as a ”deconvolutional” layer) to increase the spatial dimensions. One of the key innovations of the U-Net architecture is the use of skip connections, which allows the model to propagate information from the contracting path to the expansive pathlatex .

These skip connections are implemented by concatenating the feature maps from a layer in the contracting path with the corresponding feature maps from the same resolution in the expansive path. Throughout the years, many deep networks models on top of the prior Unet art have been proposed. in their paper Zongwei Zhou et.al, propose the,Unet++ [12] which adds deeply-supervised encoder-decoder network where the encoder and decoder sub-networks are connected through a series of nested, dense skip pathways. Liang-Chieh et.al, [2], propose a deep convolution network using an improved version of Atrous Convolution, which is a method of dilated convolution, to achieve more accurate and effective segmentation results. The authors propose a new network structure that incorporates Atrous Convolution with dilated rates learned from data instead of being fixed. This allows the network to adapt to the content of the input image and effectively capture both fine-grained and global contextual information. The experimental results showed that the proposed method outperformed several state-of-the-art models, including the classic Unet [9], on various benchmark datasets. Another 2 recent variations of Unet are ResUNet++ [4], and DoubleU-Net,[7], both proposed by the same team. ResUNet++ architecture, aims to improve upon the original Unet by incorporating residual connections and attention mechanisms. By these improvements ResUNet++ improved the segmentation results significantly on certain datasets compared to other state-of-the-art methods. The Double Unet model consists of two different encoder -

decoder architectures, that are concatenated to each other. The first encoder pathway is the VGG19 Bernal [3] and the second one is similar to the regular Unet art O.Ronnenberger et al. [9]. In this model there are skip connections between the first component to the second one, but unlike our model, these connections are only from encoder to decoder. In addition there is a concatanation action between the output of the first component, to the last one. The Double Unet was nominated for the best paper award (IEEE CBMS). in addition it outperformed the regular Unet architecture on several datasets, and other improved versions of the Unet architecture. We selected this model as a benchmark due to its impressive results, relational recognition, and the fact that unlike other published models, its code is accessible on github, so we could train and test it simply.

3 Proposed Architecture

Figure 2.1. shows a high-level overview of the suggested architecture. We take advantage of the skip connections that exists in the prior Unet art and propose an architecture called MGunet which is a multi-Grid cnn consists of n concatenated interconnected U-net components. In this model, for each middle U-net component, at any encoding level there is a skip connection from the decoder at the same level, in the previous Unet component. The main advantage of this architecture is that already in the encoder (starting with the 2nd Unet unit) each layer is aware of all the image, both on local (pixel) and global scales. This is particularly useful for classifiers, but also appears advantageous for segmentation tasks. We expect to get more comprehensive representation of information at every scale of the image that will lead to better results over the regular Unet architecture. In addition we added a classification head to this model. The classifier extracts information from the bottleneck layer in the last U-net, then processes it through a series of operations, including global max pooling and fully connected layers, before applying a softmax activation function. The output of this process is a normalized probability vector, which represents the predicted probability of each class label for a given input sample. All MGunet components are built the same, the convolution network used for this model is more narrow than the regular UNet, which mean that there are less filters in each level in comparison to the corresponding level in the Unet model. By doing so, the parameters amount ratio between Unet and MGunet is in a reasonable range.

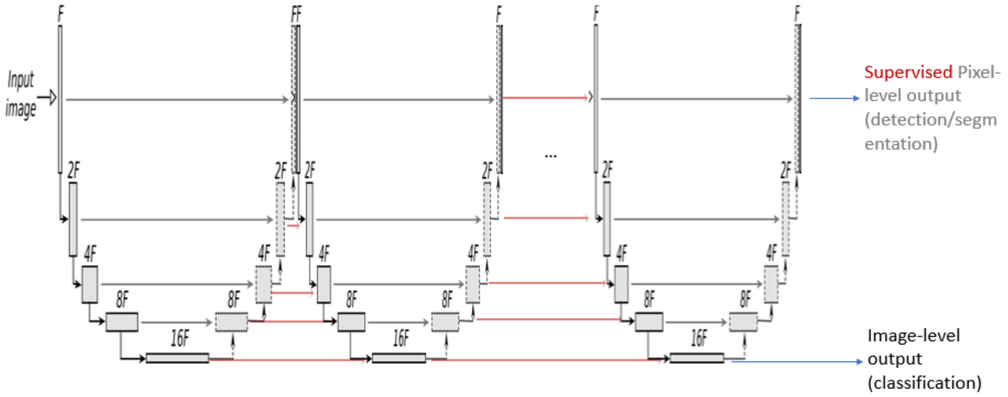

Input image
F
2F
4F
8F
16F
FF
...
Supervised Pixel-level output (detection/segmentation)
Image-level output (classification)

Figure 2.1: A high level overview of the MGUnet architecture.

4 experiment

In this section, we describe the datasets, evaluation metrics, experimental setup and configuration, and data augmentation techniques used to enhance the training process.

4.1 Datasets

Since we wanted to compare our model with the DoubleUnet model [18], and the prior Unet art,we used 3 publicly available datasets from the medical domain where these model demonstrated high accuracy and robust performance.

- CVC-ClinicDB - an open-access dataset that consists of 612 images with a resolution of 384x288, taken from 31 colonoscopy sequences. It is widely used for medical image segmentation tasks, specifically for the detection of polyps in colonoscopy videos

- Ham1000 - a large collection of multi-source dermatoscopic images of common pigmented skin lesions. The training set includes lesion from patients referred to a tertiary European referral center specialized for early detection of melanoma in high-risk groups. This group of patients often have a high number of pigmented lesions and are at a higher risk of developing malignant melanoma [10].

- 2018 Data Science Bowl Hamilton [5] - from the 2018 Data Science Bowl challenge, (**add link**) competition a nuclei segmentation. The images in this dataset form a diverse collection of biological images collectively containing tens of thousands of nuclei and the task is to segment all the nucleus in each picture [1].

4.2 Experiment setup and configuration

All models were implemented using the PyTorch framework and the implementation can be found on our repository [1]. The hardware utilized for our experiments was an Nvidia GTX1060, however, it should be noted that the amount of resources allocated may vary from experiment to experiment as we work on a remote terminal and resources are allocated per job, thus, not necessarily identical each time.

For all datasets, The data was randomly split into training, validation, and testing sets with a proportion of 70%, 15%, and 15% respectively. Due to limitations in memory, we had to reduce the image size to 128x128 for the HAM10000, and pascal datasets. We resized the rest of the datasets to 256x256. During training, we employed a loss function combined of (equally) the cross-entropy loss function (pixel-wise), and diceloss function, and used Adam optimizer to minimize this combined loss function during training. The batch size was 3 for all datasets, with the exception of the HAM10000, where the batch size was 20, and all models were trained for 300 epochs. Furthermore, we utilized a learning rate of 0.2e-3 for all training processes.

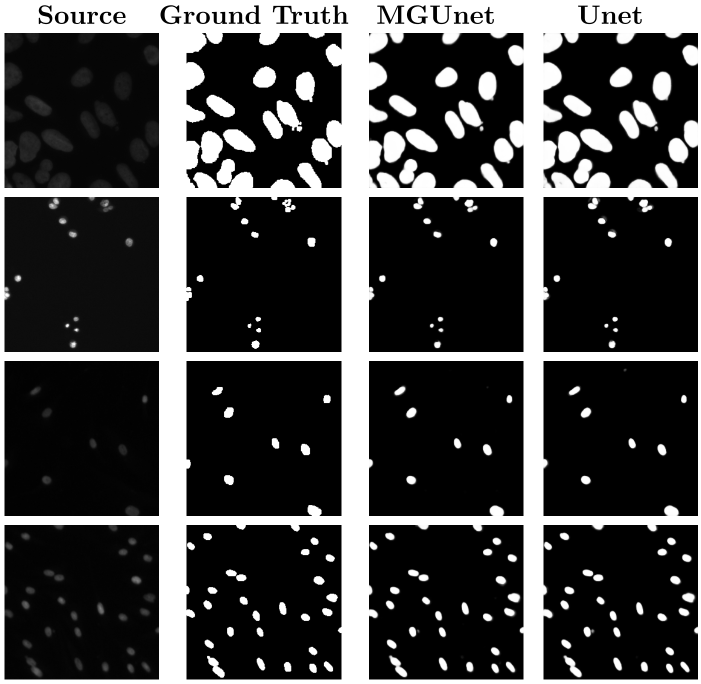
4.3 Data Augmentation

Due to the limited availability of annotated samples, applying data augmentation on the training data,allows for the network to be exposed to a diverse set of variations and deformations. One key technique that we employed in our training process is the use of color deformation. By randomly changing the brightness, contrast, saturation and hue of an image using the ColorJitter function from the torch transforms library, we were able to effectively augment the training data and improve the network's performance. Additionally, we also employed randomly various other augmentation techniques such as center crop, random rotation, and transpose, which further increased the diversity of the training data.

5 Results

In this section, we present the proposed model results from the examination of four datasets that is mentioned in 4.1. We trained 2 versions of MGunet model, defined as Narrow MGunet and Wide MGunet , both models differ from each other in the number of filters as explained in the introduction. Consequently, the number of parameters in the Wide MGunet is greater than the number of parameters in the Narrow MGunet by approximately 28.5%. In all experiments Both MGunet model versions were compared with the state of the art model double unet as SOTA, and 2 Unet models -Wide and Narrow, where the Wide Unet has the same convolution depth as in the Unet model,proposed by Olaf Ronneberger, Philipp Fischer, and Thomas Brox in their paper ([9]). of the. The performance is measured using the Intersection over Union (IOU) metric and the Dice loss function across all datasets. For each dataset we provide both the quantitative results of each model on the dataset, as well as the qualitative results, to demonstrate the effectiveness of the MGunet.

5.1 Comparison on 2018 Data Science Bowl challenge dataset

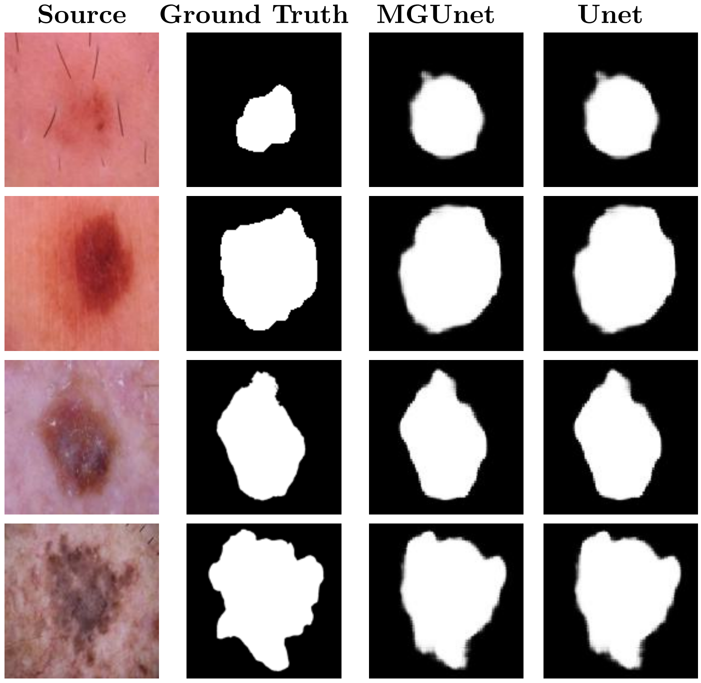


Model name	Number of parameters	IOU	Dice
Narrow MGUnet	27M	0.855 ± 0.004	0.917 ± 0.002
MG Wide	36M	0.849 ± 0.004	0.912 ± 0.001
Unet Wide	31M	0.845 ± 0.005	0.913 ± 0.002
Double Unet	29M	0.844 ± 0.005	0.913 ± 0.001
Narrow Unet	18M	0.84 ± 0.005	0.909 ± 0.002

The quantitative results presented in table shows that the Narrow MGunet achieved a Diceloss of 0.0837 and IOU of 0.145. Observing that the Narrow MGunet outperforms the DoubleU-Net by a margin of approximately 7.58% in IOU and by a margin of approximately 4.65% in Diceloss. In addition the Narrow MGunet also outperforms the rest of the models by margins of approximately 4.13%, 6.89%, 10.34% in IOU and by by margins of approximately 6.21%, 4.42%, 8.96%.

It can be seen also that due to its limited capacity, the Narrow Unet is Underperforming on this dataset. in contrast, we can see that both Narrow MGUnet, and Wide Unet models were found highly effective on this model.

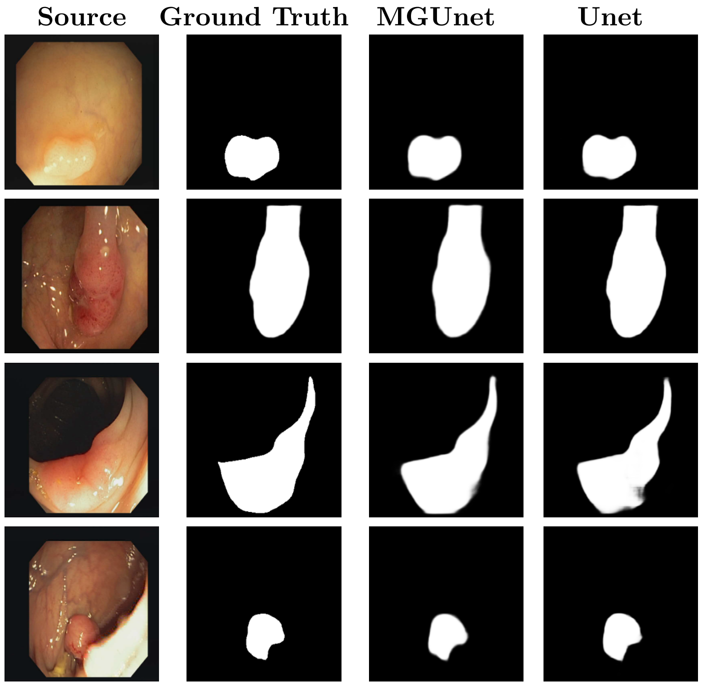
5.2 Ham10000



Model name	Number of parameters	IOU	Dice
Narrow MGUnet	27M	0.8549 ± 0.0044	0.9163 ± 0.0014
MG Wide	36M	0.8487 ± 0.0048	0.9112 ± 0.0016
Unet Wide	31M	0.8450 ± 0.0050	0.9126 ± 0.0016
Double Unet	29M	0.8437 ± 0.0051	0.9123 ± 0.0016
Narrow Unet	18M	0.8392 ± 0.0054	0.9087 ± 0.0017

Table 5.2 and Figure 5.2 present the quantitative and qualitative results obtained on the HAM10000 dataset. The Narrow MGUnet produced the best results in both measurments with margins of 0.78% in IOU, and 0.38% in Dice Loss. It can be seen from the quantitative results that all models are highly effective on this dataset and that the difference in terms of loss is very slight, especially between Wide Unet and the Narrow MGUnet. The qualitative results show that both Narrow MGUnet and Unet Wide struggle to accurately predict the borders of the ground truth, when it has irregular, jagged edges (as in lines 1,2,4 in Fig 5.2). The second line in Table 5.2 affirms that when the cell's borders in the ground truth consists of primarily staright lines, the predicted result is better.

5.3 Comparison on CVC-Clinic DB



Model name	Number of parameters	IOU	Dice
Narrow MGUnet	27M	0.9479 ± 0.1407	0.9096 ± 0.1296
MG Wide	36M	0.9386 ± 0.138	0.8982 ± 0.1264
Unet Wide	31M	0.9504 ± 0.1415	0.9206 ± 0.1327
Double Unet	29M	0.9794 ± 0.1502	0.962 ± 0.1449

Table 5.3 and Figure 5.3 present the quantitative and qualitative results obtained on the CVC-Clinic DB dataset. From the above table, we can observe that all models achieved very high IOU and diceloss scores. The qualitative results affirm the high effectiveness of the models as well. It can be seen also that the Narrow MGUnet is capable of producing better segmentation for a more challenging image (as the 3rd picture). For simpler pictures it seems that indeed the Unet produces slightly better segmentation than the Narrow MGunet. The overall qualitative results show that our proposed model performs well for different polyp frames image segmentation datasets.

6 conclusions and future research

In this paper we have proposed a new architecture, built on top of the Unet architecture. The MGunet consists of 4 concatenated interconnected U-Net components. We adjusted the model to be used for classification tasks as well by adding a classification head on top of the bottleneck output.ADD classification conclusions. On 2 out of num of datasets the performance of the proposed Narrow MGunet is significantly better in comparison to the SOTA DoubleU-Net model, and both Unet architectures we considered as baseline, while the number of parameters of the model is lower than the other (besides the Narrow Unet that obtained the worst results on 2 out of 3 datasets). Moreover the structure of this model allows much more flexibility in hyperparameters tuning. We believe that by concatenating Unet components with different blocks,the performance of this model can get better. This type of hypertuning may decrease the number of parameters in the model as well. In addition the fact that there are multiple bottlenecks can be leveraged to optimize data transmission, and transformation processes. From segmentation perspective, we propose for future research to add a modified MGUnet model, which is composed of 2 concatenated modified Unet models arranged in a "W" shape. This model would differ from previous design in that it would not use full deconvolution before the second set of convolution layers and would require adjustments to the skip layers method to align with this new model architecture. From classification perspective, another improvement in results may be obtained by building a classification head on top of each bottleneck in the network, and returning a weighted distribution classification vector out of all the classification heads outputs.

References

[1] Git repository. URL https://github.com/floccinauc/MGUnet.git.

[2] L.-C. C. G. P. F. S. H. Adam. Rethinking atrous convolution for semantic image segmentation. *ARXIV*, 2017.

[3] S. F. J. F.-E. G. G. D. R. C. . V. F. Bernal, J. Wm-dova maps for accurate polyp highlighting in colonoscopy: . computerized medical imaging and graphic, 2015. Computerized Medical Imaging and Graphic, https://polyp.grand-challenge.org/CVCClinicDB/.

[4] M. A. R. D. J. T. d. L. P. H. . D. J. ebesh Jha‡, Pia H. Smedsrud†§. Resunet++: An advanced architecture for medical image segmentation. *IEEE*, 2019.

[5] B. A. Hamilton. Broad bioimage. URL http://www.datasciencebowl.com/.

[6] N. Ibtehaz1 and . M. Sohel Rahman1. Multiresunet : Rethinking the u-net architecture for multimodal biomedical image segmentation. *ARXIV*, 2019.

[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. pages 558–564, 2020.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[9] O.Ronnenberger, P.Fischer, and T.Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[10] P. Tschandl. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. URL https://www.nature.com/articles/sdata2018161.

[11] C. H. M. D. O. R. Z. Xuebin Qin, Zichen Zhang and C. Martin Jagersand University of Alberta. $\mathrm{U}^2 - net : Goingdeeperwithnestedu - structureforsalientobjectdetection.ScienceDirect$, 2022.

[12] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation, 2018. URL https://arxiv.org/abs/1807.10165.