EasyChair Preprint

№ 14733

SUDS: a Simplified U-Net Architecture with Depth-Wise Separable Convolutions

Vlad-Constantin Ionete and Cosmin Marsavina

September 6, 2024

SUDS: A Simplified U-Net Architecture with Depth-wise Separable Convolutions

Vlad-Constantin Ionete
Department of Computers and Information Technology
University Politehnica of Timisoara, Romania
vlad.ionete@student.upt.ro

Cosmin Marsavina
Department of Computers and Information Technology
University Politehnica of Timisoara, Romania
cosmin.marsavina@upt.ro

***Abstract*—Medical image segmentation is one of the most important topics in the field of computer vision and plays a crucial role in computer-aided diagnosis. U-Net paved the way for a series of variants that took advantage of the key characteristics of this network. In this article, several features proposed in different variants of U-Net are adapted and experimented upon to create a new architecture that maintains the idea of a U-shaped structure. The proposed architecture takes advantage of the efficient depth-wise separable convolution, but with a twist. Instead of using the pointwise convolution as the last step in the depth-wise separable convolution, it utilizes the so-called Ghost Module. This results in a highly efficient network with a reduced complexity, that still has excellent segmentation performance. We compared SUDS with U-Net and its variants across multiple segmentation tasks from two categories, skin lesion segmentation and colonscopy segmentation. Experiments demonstrate that SUDS has similar segmentation accuracy compared to the other networks, while the number of parameters and floating-point operations are greatly reduced.**

***Index Terms*—computer vision, medical image segmentation, U-Net**

I. Introduction

Medical imaging devices such as MRI, CT, X-ray, and colonoscopy probing equipment are crucial for clinical diagnosis. A key technique in this area is medical image segmentation, which heavily relies on advancements in Convolutional Neural Networks (CNNs). Starting from the pioneering CNN, LeNet by LeCun *et al.* [1], there have been significant developments with networks such as AlexNet by Krizhevsky *et al.* [2], VggNet by Simonyan and Zisserman [3], and GoogleNet by *Szegedy et al.* [4]; these laid the groundwork for later innovations, such as ResNet [5] and EfficientNet [6].

In recent years, CNNs have excelled in classifying each pixel of an image, thereby enhancing image segmentation capabilities as noted by Wolterink *et al.* [7] in 2017. A landmark development in medical image segmentation was the introduction of U-Net [8], which has become foundational for many biomedical segmentation tasks. Following U-Net, various adaptations have emerged, thus enhancing its capabilities. For instance, TransUNet [9] combines Transformers with U-Net to boost segmentation accuracy. The 3D U-Net model adapts the U-Net architecture for 3D imaging, thereby improving volumetric segmentation with fewer annotations. Attention U-Net [10] integrates an attention mechanism, particularly aiding in the segmentation of the pancreas.

It is imperative to also acknowledge other influential networks in image segmentation such as DeepLab [11], which utilizes atrous convolution and CRFs for high-resolution segmentation, and Capsule Networks [12], which offer unique methodologies in this field. More details on U-Net and its significant variants will be discussed in subsequent sections.

II. Related Work

The advancement of semantic segmentation techniques, particularly in the realm of medical imaging, has seen a remarkable trajectory spurred significantly by the advent and continuous refinement of Convolutional Neural Networks (CNNs). A pivotal moment in this journey was the introduction of the U-Net architecture by Ronneberger *et al.*, which was designed specifically for biomedical image segmentation. This architecture, with its innovative contracting and expanding paths, not only set new benchmarks for segmentation accuracy, but also introduced a novel way of utilizing data augmentation for training on limited datasets.

A. Evolution of U-Net: from foundations to enhanced variants

The foundational U-Net architecture laid the groundwork for subsequent research aimed at addressing the unique challenges in medical image segmentation. Its design was particularly revolutionary for allowing end-to-end training with very few images, leveraging data augmentation to significantly enhance the model's performance. This breakthrough has catalyzed a wave of innovations, seeking to adapt and refine the U-Net architecture for broader applications and improved efficiency.

Hasan and Linte's development of U-NetPlus [13] is a notable example of these efforts. By integrating a pre-trained encoder and re-envisioning the decoder through nearest-neighbor (NN) interpolation for upsampling, U-NetPlus is aimed at enhancing segmentation performance in surgical instrument detection. This adaptation was particularly geared towards overcoming challenges inherent in robotic-assisted surgeries (such as occlusions or varying illumination), thus demonstrating the potential of deep learning in complex, dynamic environments.

B. Computational efficiency in semantic segmentation

A critical aspect in the evolution of segmentation models has been the emphasis on computational efficiency. The introduc-

tion of depth-wise separable convolutions, as explored in various studies, represents a significant stride towards reducing the computational burden of segmentation tasks. These methods have proven effective in decreasing the number of parameters and the computational requirements, thereby facilitating faster processing times without compromising accuracy. This is crucial for applications requiring real-time analysis, such as intraoperative surgical assistance.

C. Towards adaptive and resource-efficient architectures

The integration of depth-wise separable convolutions within U-Net architectures, as exemplified in the study performed in "An Image Deblurring Method Using Improved U-Net Model" [14] marks a significant moment in the pursuit of more efficient and effective semantic segmentation models, particularly for medical imaging applications. This approach is complemented by the recent development of Half-UNet, a model that embodies the quest for balancing model complexity with computational efficiency, thereby addressing some of the most pressing challenges in medical image analysis.

D. Depth-wise separable convolutions and U-Net enhancements

Depth-wise separable convolutions offer a computationally efficient alternative to standard convolutions by decoupling the filtering and combining the phases of the convolution process. This results in a significant reduction in both the number of parameters and the computational cost whilst also addressing the vanishing gradient problem, thereby making it an ideal choice for tasks that require real-time processing or that operate within resource-constrained environments. The "An Image Deblurring Method Using Improved U-Net Model" article demonstrates how integrating these convolutions into a U-Net framework can lead to substantial improvements in processing efficiency without sacrificing the model's ability to effectively perform image deblurring tasks, a common requirement for enhancing the usability of medical imagery.

E. Half-UNet: streamlining U-Net for efficiency

In parallel with these developments, the concept of Half-UNet [15] has emerged as another innovative approach to streamline the U-Net architecture. This variant seeks to maintain the essence of U-Net's design (namely its ability to capture detailed contextual information while ensuring precise localization) within a more compact and computationally less demanding framework. By rethinking the architecture to include fewer convolutional layers and by optimizing the network's path flows, Half-UNet aims to offer a solution that is not only adept at handling segmentation tasks but also markedly more efficient in terms of computational resource utilization.

F. Bridging the gap between theory and application

Both the integration of depth-wise separable convolutions into U-Net and the conceptualization of Half-UNet underscore a broader trend in medical image analysis, namely the drive towards creating models that are not only powerful in terms of segmentation accuracy but also optimized for speed and efficiency. This is particularly crucial in clinical settings, where the ability to quickly and accurately process images can significantly impact diagnostic workflows and patient outcomes.

The advancements brought forward by these models open new avenues for research, particularly in exploring further optimizations that can reduce computational load without compromising accuracy. The continuous refinement of U-Net and its variants reflects a tendency towards creating more adaptive and resource-efficient architectures. Such advancements underscore the ongoing efforts to balance the trade-offs between model complexity, computational demands, and segmentation performance.

III. Proposed Model

A. Model structure

The SUDS-Net architecture represents a modern adaptation of the U-Net model, specifically designed for semantic image segmentation with a keen focus on enhancing computational efficiency. As shown in Figure 1, the proposed architecture is symmetric, featuring an encoder-decoder framework that includes Residual Depth-wise Separable Blocks and the innovative Ghost Module. The network is characterized by its depth-wise separable convolutions for efficient feature extraction and the Ghost Module that further optimizes the computational cost by generating additional feature maps through fewer computations. The depth-wise separable convolutions and Ghost Module are not present in the architectural scheme because they are integrated in the Residual Depth-wise Separable Blocks, but will be presented later on. The activation function used is ReLU and the kernel size is 3×3.

The encoder accomplishes the extraction of image information in two stages. In the first stage the model processes the input, which in this example is assumed to be an RGB image. This stage is also the first encoder block; the block applies a Residual Depth-wise Separable Convolution (with potential Ghost Module enhancements) to increase the channel depth from 3 to 64 while capturing spatial features, effectively preparing the feature maps for deeper processing. The block is followed by a Max Pooling layer to reduce spatial dimensions by half, thus enhancing the model's ability to capture higher-level features at reduced resolutions. The second stage or second encoder block is similar to the first one; this block further processes the feature maps, doubling the channel depth to 128 (from 64), and is also followed by Max Pooling. The bridge is tasked with synthesizing the highest-level features extracted by the encoder. It concentrates on capturing the essence of the input image's content, which is crucial for the decoder to generate a precise segmentation map.

Unlike traditional bottleneck layers that might simply compress features, the bridge in SUDS-Net also enhances them through depth-wise separable convolutions and the Ghost Module. This ensures that the transition to the decoder is not merely a passing of information but an active enhancement of it. In this specific architecture, the bridge can be

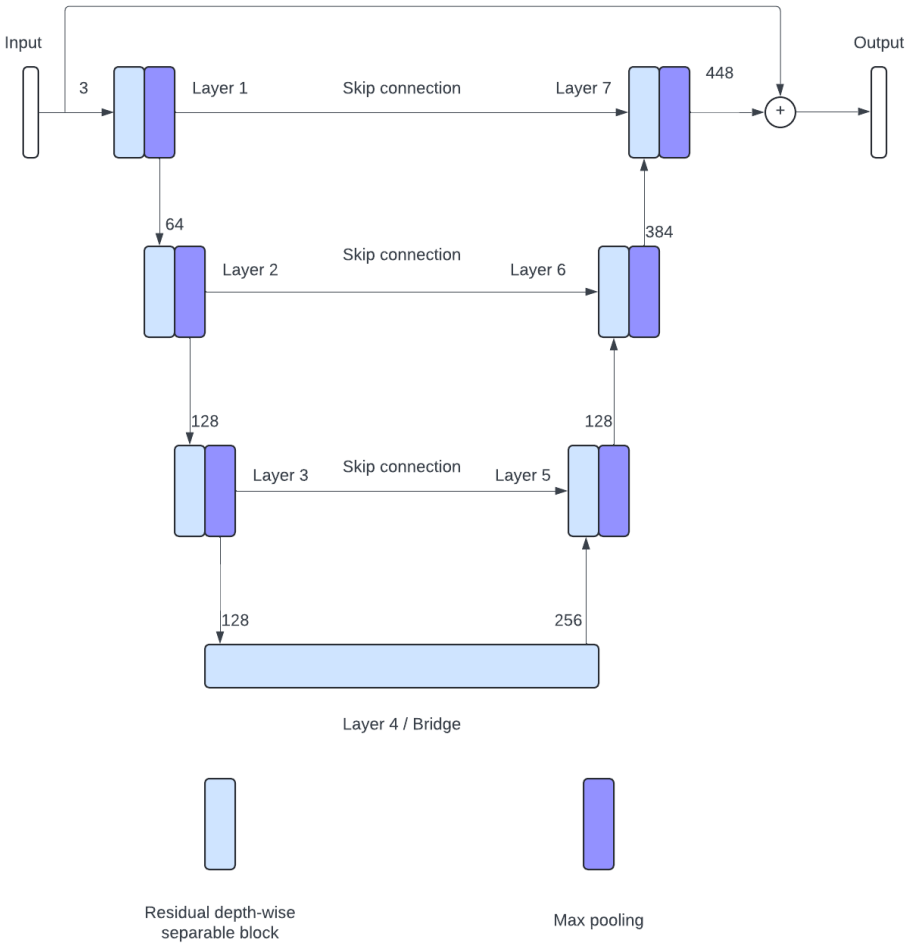


Fig. 1: The architecture of the model

considered an extension of the encoder's feature extraction process, albeit with a distinct purpose of synthesizing the highest-level features before the decoding process begins. The bridge increases the depth from 128 channels to 256. The decoder reconstructs the segmentation map from the encoded feature maps, progressively upscaling and merging features from the encoder through skip connections. This happens in two stages; the first decoder stage begins with the bridge output (assumed to be 256 channels based on the bridge's final output). It upsamples the feature maps using transposed convolution, reducing the channel depth to 128 to increase spatial resolution. It concatenates these upsampled features with the corresponding encoder output from the second stage (128 channels), effectively doubling the channel depth temporarily before processing. After the concatenation the output is 384 channels. Everything is processed through a Residual Depth-wise Separable Block with Ghost Module. The second stage follows a similar pattern to the first one, further upsampling and processing the feature maps. It upsamples the input channels (384) from the previous decoder block and concatenates them with the corresponding output channels from the encoder block (64), resulting in 448 channels. The operations are performed by a Residual Depth-wise Separable Block with Ghost Module that further refines the segmentation map, upsampling and fusing features as before, in preparation for the final segmentation output.

The final layer maps the decoder output to the desired number of classes for segmentation, thus producing the final segmentation map. It is represented by a convolution layer that takes as input the output from the last decoder block (448) and it outputs the number of classes for the model.

B. Important features

The Simplified U-Net with Residual Depth-wise Separable Blocks (SUDS-Net) introduces a refined architecture for semantic image segmentation, drawing inspiration from the foundational elements of Half-UNet and advancements in depth-wise separable convolutions (as demonstrated in the Improved U-Net Model for deblurring). Our model leverages these concepts to enhance computational efficiency while aiming to maintain or improve the segmentation accuracy of traditional U-Net models.

1) Encoder-decoder framework: At the heart of SUDS-Net lies an encoder-decoder framework enhanced with depth-wise separable convolutions. This design choice is motivated by the need to reduce computational complexity and the model's parameter count, a principle that is central to the design of Half-UNet.

2) Depth-wise convolution: Applies a single filter per input channel, significantly reducing the computational complexity compared to standard convolutions, as shown in Figure 2.

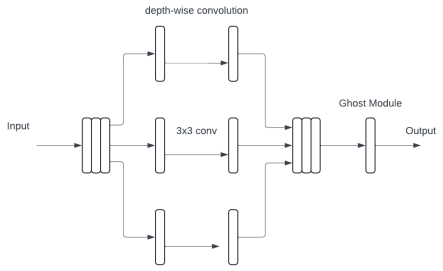


Fig. 2: The depth-wise separable convolution structure

For an input feature map X with $[H, W, C_{in}]$, the depth-wise convolution operation can be defined as:

$$DW(X) = X * K_{dw} \tag{1}$$

3) Pointwise convolution replaced by Ghost Module: Traditionally, the output of the depth-wise convolution is further processed by a pointwise convolution to combine the channel-wise features. In SUDS-Net, this step is innovatively replaced by a Ghost Module.

4) The Ghost Module: Is a novel component introduced in Half-UNet. In our architecture the Ghost Module replaces the pointwise convolution, aiming to generate more feature maps through fewer computations. The module operates by generating "ghost" feature maps through cheap operations, thereby reducing the number of parameters and computational demands.

$$GM(P) = [P * K_{\text{primary}}, \phi(P * K_{\text{cheap}})] \tag{2}$$

where P is $DW(X)$, K_{primary} is the kernel for the primary convolution within the Ghost Module, K_{cheap} represents the kernel for cheaper operations to generate additional feature maps, and ϕ is a cheap operation (such as a depth-wise convolution). The output of GM is the concatenation of the primary convolution output with the transformed cheap operation output.

5) Residual Depth-wise Separable Block: Allows the input to bypass the depth-wise (as seen in Figure 3) and pointwise (or Ghost Module) convolutions, adding it directly to the block's output to facilitate gradient flow and mitigate the vanishing gradient problem. If X is the input and $F(X)$ represents the combined operation of depth-wise convolution followed by the Ghost Module, then the Residual Depth-wise Separable Block output can be represented as:

$$R(X) = F(X) + X \tag{3}$$

assuming the dimensions of $F(X)$ and X are compatible for addition. If dimensionality needs adjustment, a 1×1 convolution C_{1x1} is applied to X before addition:

$$R(X) = F(X) + C_{1x1}(X) \tag{4}$$

Combining these components, the operation of a Residual Depth-wise Separable Block with a Ghost Module, denoted as $RDSB_{\text{GM}}(X)$, can be formalized as:

$$RDSB_{\text{GM}}(X) = H(X) + \begin{cases} X, & \text{if } C_{in} = C_{\text{out}} \\ C_{1\times1}(X), & \text{otherwise} \end{cases} \tag{5}$$

where C_{in} and C_{out} are the input and output channel sizes, respectively and

$$H(X) = ReLU(GM(DW(X))) \tag{6}$$

While Half-UNet focuses on streamlining the U-Net architecture for general efficiency, SUDS-Net extends these principles specifically towards enhancing the model's performance in segmentation tasks that benefit from reduced computational complexity.

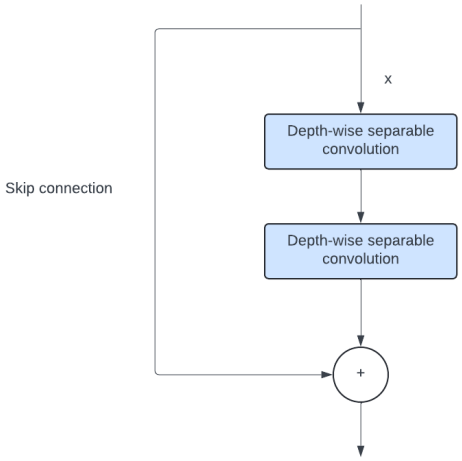


Fig. 3: Residual Depth-wise Separable Block

C. Performance

1) Number of parameters: The number of parameters in a convolutional layer is determined by the size of its kernels and the number of input and output channels. For a standard convolutional layer with kernel size $K_w \times K_h$, input channels C_{in} and output channels C_{out}, the numbers of parameters is given by:

$$\text{Params} = (K_W \cdot K_h \cdot C_{\text{in}} + 1) \cdot C_{\text{out}} \tag{7}$$

The "+1" accounts for the bias term for each output channel, which can be omitted if bias is set to False. For a depth-wise convolutional layer, since there is only one filter per input channel and no mixing of channels, the number of parameters is significantly less:

$$\text{Params}_{\text{depth-wise}} = K_w \cdot K_h \cdot C_{\text{in}} \tag{8}$$

For the Ghost Module with primary and cheap operations the formula for parameters is the following:

$$\text{Params}_{\text{Ghost}} = \text{Params}_{\text{primary}} + \text{Params}_{\text{cheap}} \tag{9}$$

where $\text{Params}_{\text{primary}}$ is the number of parameters for the primary convolutions in the Ghost Module while $\text{Params}_{\text{cheap}}$ is the number of parameters for cheap operations (like depth-wise convolutions) in the same module. For a residual block, the number of parameters is the sum of the parameters from the depth-wise separable convolutions with a Ghost Module and those from the 1×1 convolution if used to match the channel dimensions.

$$\text{Params}_{\text{residual}} = \text{Params}_{\text{depth}} + \text{Params}_{\text{Ghost}} + \text{Params}_{\text{short}} \tag{10}$$

where $\text{Params}_{\text{short}}$ (short stands for shortcut) is calculated similarly to a standard convolution:

$$\text{Params}_{\text{short}} = C_{\text{in}} \cdot C_{\text{out}} \quad \text{if } C_{\text{in}} \neq C_{\text{out}} \text{ or stride} \neq 1 \tag{11}$$

2) FLOPs: The FLOPs for a layer are calculated by considering the number of multiplications and additions for each operation in the forward pass. For a standard convolutional layer, the FLOPs can be approximated as:

$$\text{FLOPs} = (2 \cdot K_w \cdot K_h \cdot C_{\text{in}} - 1) \cdot H_{\text{out}} \cdot W_{\text{out}} \cdot C_{\text{out}} \tag{12}$$

Here H_{out} and W_{out} represent the height and width of the output feature map and the factor of 2 accounts for both multiplication and addition operations in the convolution. We subtract 1 if there is no bias.

For the depth-wise convolutions, FLOPs are computed as:

$$FLOPs_{\text{depth-wise}} = K_w \cdot K_h \cdot C_{\text{in}} \cdot H_{\text{out}} \cdot W_{\text{out}} \tag{13}$$

For the Ghost Module, FLOPs are calculated as:

$$\text{FLOPs}_{\text{Ghost}} = \text{FLOPs}_{\text{primary}} + \text{FLOPs}_{\text{cheap}} \tag{14}$$

For the primary operation, the formula for FLOPs is:

$$FLOPs_{\text{primary}} = (2 \cdot K^2 \cdot C_{\text{in}} - 1) \cdot H_{\text{out}} \cdot W_{\text{out}} \cdot C_{\text{primary}} \tag{15}$$

For the cheap operation, the formula is:

$$FLOPs_{\text{cheap}} = (2 \cdot K_{\text{dw}}^2 \cdot C_{\text{primary}} - 1) \cdot H_{\text{out}} \cdot W_{\text{out}} \cdot C_{\text{cheap}} \tag{16}$$

IV. Datasets, equipment and technologies used

We validate our network model using two public datasets, as shown in Table I. Even though there are relatively few images in each dataset, the model works better without performing data augmentation. This will be shown in the following section which addresses results.

The ISIC 2018 ([16], [17]) dataset is a comprehensive resource for the development and benchmarking of machine learning models in the domain of dermoscopy image analysis, particularly for tasks related to skin lesions. Released by the International Skin Imaging Collaboration (ISIC), the dataset has been instrumental in several challenges aimed at advancing research in melanoma detection. The ISIC 2018 dataset is part of a series of annual challenges that provide a platform for participants to test their models against a standard benchmark. The dataset features a large-scale collection of dermoscopy images that can be used for different tasks, including lesion segmentation (Task 1) and lesion attribute detection (Task 2); our study focuses on Task 1, the lesion segmentation task. It includes 2594 dermoscopic lesion images, each paired with a corresponding binary mask indicating the primary skin lesion's location. The input images are in JPEG format and have a unique 7-digit identifier, while the response data are binary mask images in PNG format. The goal of Task 1 is to submit automated predictions of lesion segmentation boundaries within the dermoscopic images. The challenge emphasizes that each lesion image contains exactly one primary lesion and any other pigmented regions or markings should be disregarded.

For training, participants have access to the images and the ground truth data. The ground truth segmentations were reviewed and curated by dermatologists and created using various methods, including fully-automated algorithms, semi-automated techniques, and manual tracing. The data is split into 2594 images for the training set, 100 for the validation set and 1000 images for the test set.

The CVC-ClinicDB [18] dataset is a collection of 612 high-resolution images (384x288 pixels) from 31 colonoscopy sequences, specifically curated for medical image segmentation tasks. This dataset is particularly focused on the detection of polyps in colonoscopy videos. Each image in the dataset comes with a corresponding annotation mask that delineates the polyp, providing essential ground truth for segmentation algorithms. Researchers and developers commonly use the CVC-ClinicDB dataset to develop and test algorithms for automated polyp detection, which is a crucial task in the early diagnosis and treatment of colorectal cancer. The dataset serves as a benchmark in the field of medical image analysis, allowing for the comparison of different segmentation models in terms of performance.

Given that the dataset contains images with pixel-level semantic segmentation annotations, it is well-suited for deep learning models designed to understand and interpret visual data within the medical domain. The dataset provides a realistic challenge for models due to the variability in polyp appearance and size, as well as the complexities of the internal structures visible through colonoscopy imagery. The data does not come split into training, validation and test data. For our experiments, we split the data as follows: 427 images for training, 123 for validation and 62 for testing. This dataset is freely available on the Dataset Ninja website.

The training of the model was conducted on a gaming laptop equipped with a NVIDIA RTX 3070 GPU, featuring 8GB of VRAM. The upcoming section will demonstrate that, despite utilizing a less powerful GPU, our architecture achieves good results (comparable to established U-Net variants). It is anticipated that with superior hardware the model should exhibit even more pronounced improvements over the baseline. The model was developed with a minimalist setup (starting from the original U-Net) using Python, with PyTorch as the only framework employed.

TABLE I: Datasets and their characteristics

Dataset	Images	Input size	Provider
ISIC	2594	1022 x 767	ISIC
CVC	612	384 x 288	CVC

V. Results

A. Experimental study details

As detailed in the preceding discussion, our architecture underwent training on two distinct datasets. To ensure a fair evaluation, all the networks were trained using Adaptive Moment Estimation (Adam) across differing epochs (25 for the ISIC dataset and 60 for CVC) with an initial learning rate set to 0.001. Training involved the use of mini-batches comprised of 4 images each and the Cross-Entropy Loss Function was utilized as the loss criterion.

The datasets prepared for the comparison with the other networks include data augmentation and L2 regularization was applied across all models barring our own architecture. Interestingly, subsequent analysis revealed that data augmentation and regularization, while commonly employed, may not be necessary and could potentially impair our model's performance. This observation will be further elaborated upon in the following sections, thus highlighting our model's unique approach and its implications for overall effectiveness.

B. Evaluation indicators

In this study, the segmentation performance is evaluated through three metrics: Intersection over Union, sensitivity and specificity.

Intersection over Union (IoU) is a metric used to evaluate the accuracy of an object detector on a particular dataset. It is often utilized in computer vision tasks to measure how well a predicted bounding box (or segmentation mask) overlaps with the ground truth bounding box (or mask). IoU is a simple yet powerful metric that quantifies the size of the overlap between two shapes. It is defined as the size of the intersection divided by the size of the union of the two shapes. Given two sets A and B where A represents the predicted bounding box or segmentation mask, and B represents the ground truth bounding box or mask, the IoU is defined as:

$$\text{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{17}$$

where $|A \cap B|$ denotes the area of overlap between the two sets and $|A \cup B|$ denotes the combined area of A and B (including their overlap).

Sensitivity, also known as recall or the true positive rate, is a measure of the proportion of actual positive cases that are correctly identified by the model. It represent a key metric in many fields, especially in medical diagnostics where it is important to identify as many true cases of a condition as possible. Mathematically, sensitivity is defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{18}$$

Specificity, also known as the true negative rate, measures the proportion of actual negatives that are correctly identified.

TABLE II: Model performance comparison: parameters, FLOPs and IoU

Architecture	Params	FLOPs	IoU ISIC	IoU CVC
U-Net	31.04M	11x	0.7754	0.6534
DC-UNet	10.07M	43x	0.7812	**0.6782**
Half-UNet	0.21M	1x	0.7632	0.6631
SUDS	0.23M	1x	**0.7834**	0.6720

TABLE III: Model performance comparison: sensitivity and specificity

Architecture	Sensit ISIC	Sensit CVC	Spec ISIC	Spec CVC
U-Net	0.9211	0.8475	0.9947	0.9945
DC-UNet	0.9415	**0.8706**	0.9950	**0.9949**
Half-UNet	0.9338	0.8206	0.9934	0.9923
SUDS	**0.9432**	0.8645	**0.9952**	0.9948

It is also a critical metric in settings where the cost of false positives is high. In the context of medical testing for example, specificity indicates the likelihood that a test can correctly identify the individuals without a disease when they are truly disease-free. This metric helps minimize the risk of incorrectly diagnosing healthy patients as sick. Mathematically, specificity is defined as:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{19}$$

C. Experimental results

In our comparative study, we scrutinize the efficacy of the SUDS framework against U-Net and its derivatives, particularly focusing on the segmentation of skin lesions and colonoscopies. We utilize the model's parameters and floating-point operations (FLOPs) as metrics to gauge the architectural complexity and computational demands. Concurrently, Intersection over Union (IoU) serves as the benchmark for segmentation prowess. As delineated in Tables II and III, the empirical data shows the superior performance of SUDS over all U-Net variants in skin lesion segmentation.

SUDS is an architecture that is less complex, as evidenced by the significantly fewer parameters (0.23M compared to 31.04M for U-Net and 10.07M for DC-UNet) and the minimal computational cost, quantified through an identical FLOPs index to Half-UNet's (1x). However, the model does not compromise on performance, achieving an IoU of **0.7834** which is marginally better than DC-UNet's 0.7812 and U-Net's 0.7754. While DC-UNet comes close to SUDS in terms of performance benchmarks, its complex architecture and longer training times represent notable drawbacks. In particular, for the CVC dataset DC-UNet edges ahead with a marginally higher IoU (0.6782 vs. 0.6720). However, this minor improvement does not fully justify its substantially greater complexity, which becomes a hindrance when dealing with datasets featuring intricate or less distinct patterns. In contrast,

SUDS exemplifies the balance between model simplicity and effective performance. It achieves this without the extensive computational demands seen in more complex models such as DC-UNet, thereby showcasing the potential of streamlined, efficient architectures in medical image segmentation tasks.

The reduction in parameters by over 97% from U-Net, without a significant drop in IoU, showcases the model's refined efficiency. This suggests that SUDS captures the essential representational power necessary for segmentation tasks without the encumbrance of computational intensity typically associated with deeper or wider network architectures. These results are indicative of our model's efficiency, where it maintains high fidelity in segmentation with substantially reduced computational overhead. SUDS embodies an optimized trade-off between network complexity and performance efficacy.

In the conducted experiments, we assessed various configurations of our model against each other with regard to their performance on the provided datasets. An intriguing observation was noted: the model devoid of data augmentation and L2 regularization demonstrated superior performance over its counterparts. Upon examining Table IV, several aspects regarding the enhanced performance of the "clean" model were observed:

- **Intersection over Union (IoU)**: a substantial improvement in IoU is evident for the clean model (0.7834) when juxtaposed against the augmented variant (0.6410). IoU is a pivotal metric in the domain of segmentation, offering a quantifiable measure of the overlap between the predicted segmentation and the ground truth annotation. The elevated IoU for the clean model suggests a better alignment with the ground truth, thereby reflecting an improved segmentation accuracy.
- **Sensitivity**: The clean model also exhibits an enhanced sensitivity score (0.9432) compared to the augmented one (0.7345), thus indicating its proficiency in correctly identifying pixels pertaining to the region of interest. This increased sensitivity is of particular importance in medical image analysis, wherein the omission of critical regions could lead to dire repercussions.
- **Specificity**: Both models achieved commendable specificity scores; however, the clean model substantially outperformed the augmented one (0.9952 vs. 0.8834). Specificity is also a vital metric, ensuring the model's precision in segmenting only the objects of interest and mitigating the misclassification of background elements. The superior specificity of the clean model proves a more discerning segmentation capability, thus potentially reducing the number of false-positive.

These findings suggest that for the specific datasets and tasks in this study, using a less complex architecture (without data augmentation and L2 regularization) may result in more accurate segmentation outcomes. It highlights the need for a nuanced approach to model selection where additional complexities such as data augmentation are carefully weighed against their actual impact on model performance.

TABLE IV: Performance metrics for ISIC models

Model	Best IoU	Sensitivity	Specificity
SUDS_CLEAN	**0.7834**	0.9432	0.9952
SUDS_DATA_AUG	0.6410	0.7345	0.8834

D. Qualitative results

The segmentation performance of SUDS, DC-UNet, Half-UNet, and U-Net is also assessed from a qualitative perspective (Figure 4). The images make it abundantly evident that the SUDS model is able to precisely determine the limits of segmented parts. When compared to the other models, SUDS offers segmentation boundaries that are exact and accurate, very close to the ground truth.

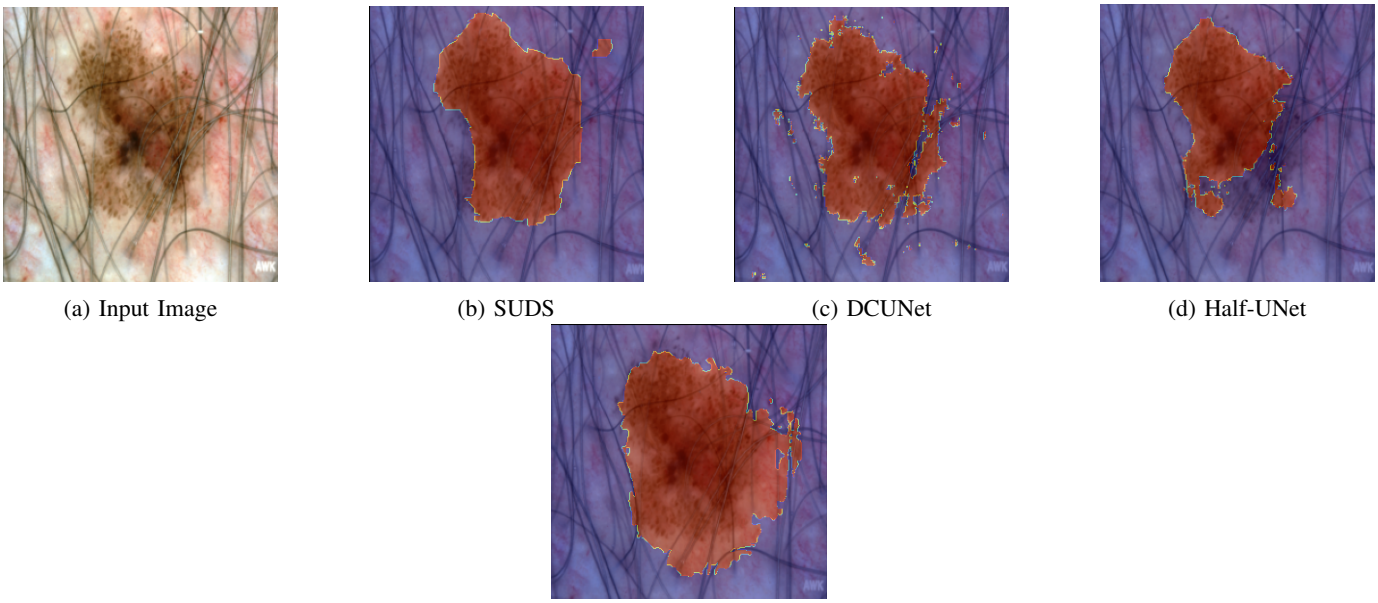
The precision and thoroughness of the outlines generated by SUDS are especially remarkable. The SUDS segmentation ground truth has a higher level of clarity and comprehensiveness, thus indicating that the model effectively captures the fine details and boundaries of the segmented areas. In striking contrast, Half-UNet, UNet and DC-UNet show less accuracy and more fragmentation in their segmentation borders.

Although computationally efficient, Half-UNet's segmentation clearly misses complex details and shows a degree of smoothness that might hide significant boundaries. U-Net surpasses Half-UNet in several aspects, but it still has issues in achieving precision in complex regions. Even though DC-UNet is more complicated, it does not consistently surpass SUDS, especially in terms of preserving the integrity of segmentation contours.

The improved efficiency of SUDS can be credited to its optimized structure which, despite having fewer parameters and lower processing demands, is able to attain a higher level of accuracy in capturing the subtle details of the target areas. The architectural efficiency results in improved segmentation quality, as illustrated in Figure 4.

VI. Discussion

Our findings indicate that SUDS outperforms U-Net and its variants in terms of segmentation efficiency. The model demonstrates versatility across different segmentation tasks, as evidenced by its performance on the two distinct datasets used in our experiments. By simplifying the U-Net architecture and incorporating proven techniques tailored to our needs, we developed a model that is both less complex and more effective. The integration of depth-wise separable convolution and the substitution of pointwise convolution with the Ghost Module significantly reduces complexity without compromising performance. As shown in Table II, SUDS operates with fewer parameters and requires less computational power (FLOPs) compared to the other models. However, it slightly underperforms on complex datasets with intricate features or suboptimal lighting conditions, as indicated by a lower Intersection over Union (IoU) compared to DC-UNet. Table IV also reveals that data augmentation negatively impacted the performance of our model in these experiments. Further



(a) Input Image (b) SUDS (c) DCUNet (d) Half-UNet

(e) UNet

Fig. 4: Segmentation output comparison for the investigated networks

improvements could be achieved by refining the model or utilizing more advanced hardware.

VII. Conclusions

In this study, we demonstrate that the SUDS architecture for medical image segmentation achieves success through strategic simplifications and the integration of specific components and techniques. SUDS maintains robust performance while reducing overall complexity by incorporating Depth-wise Separable Convolution and the Ghost Module. We validate the effectiveness of SUDS through comprehensive comparisons with U-Net and its variants, which demonstrate that the proposed architecture delivers comparable segmentation results while significantly simplifying network complexity.

Future work. We want to train SUDS on different datasets and broaden its capabilities, while also employing better hardware for training and inferring. This can go hand in hand with adapting the architecture to handle 3D medical image data, such as MRI and CT scans. Implementing 3D depth-wise separable convolutions that can maintain computational efficiency while leveraging volumetric information might be a promising area for future research. Another encouraging direction would be the integration of attention mechanisms (such as those used in Attention U-Net) to enable the model to focus more precisely on the relevant regions within the image. Attention modules could help improve segmentation accuracy, especially for challenging datasets in which the target structures exhibit significant variability. Finally, the end goal of this research would be the collaboration with clinical experts that could validate the relevance of the output provided by SUDS and its integration within clinical workflows.

References

[1] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. *Gradient-based learning applied to document recognition*, Proc. IEEE 86, 1998.

[2] Krizhevsky, A., Sutskever, I., and Hinton, G. E., *Imagenet classification with deep convolutional neural networks*, Advances in Neural Information Processing Systems, 2012.

[3] Simonyan, K., and Zisserman, A., *Very deep convolutional networks for large-scale image recognition.*

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al., *Going deeper with convolutions* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Boston, MA), 2015.

[5] He, K., Zhang, X., Ren, S., and Sun, J., *Deep residual learning for image recognition* Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2016.

[6] Mingxing Tan, Quoc V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks* International Conference on Machine Learning, 2019.

[7] Wolterink, J. M., Leiner, T., Viergever, M. A., and Igum, I. *Automatic Segmentation and Disease Classification Using Cardiac Cine MR Images*, Cham: Springer, 2017.

[8] Olaf Ronneberger, Philipp Fischer, Thomas Brox *U-Net: Convolutional Networks for Biomedical Image Segmentation*, MICCAI, 2015.

[9] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, Yuyin Zhou, *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*, 2021.

[10] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, Daniel Rueckert, *Attention U-Net: Learning Where to Look for the Pancreas*, MIDL, 2018.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*, TPAMI, 2016.

[12] Jyostna Devi Bodapati, V.N. Rohith, *ChxCapsNet: Deep capsule network with transfer learning for evaluating pneumonia in paediatric chest radiographs*, IEEE International Conference on Image Processing (ICIP), 2020.

[13] S. M. Kamrul Hasan, Cristian A. Linte, *U-NetPlus: A Modified Encoder-Decoder U-Net Architecture for Semantic and Instance Segmentation of Surgical Instrument*, IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020.

[14] Zuozheng Lian, Haizhen Wang, Qianjun Zhang, *An Image Deblurring Method Using Improved U-Net Model*, ICISCAE, 2020.

[15] Haoran L, Yifei She, Jun Tie, Shengzhou Xu, *Half-UNet: A Simplified U-Net Architecture for Medical Image Segmentation*, Frontiers in Neuroinformatics, 2022.

[16] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, Allan Halpern, *Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)*, ISIC, 2018.

[17] Tschandl, P., Rosendahl, C., Kittler, H., *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*, Sci. Data 5, 2018.

[18] Jorge Bernal and F. Javier Sánchez and Gloria Fernández-Esparrach and Debora Gil and Cristina Rodríguez and Fernando Vilariño, *A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images*, ICPR, 2014.