Enhancing Medical Image Segmentation: Optimizing Cross-Entropy Weights and Post-Processing with Autoencoders

Pranav Singh
Tandon School of Engineering
New York University
ps4364@nyu.edu

Luoyao Chen
Center for Data Science
New York University
lc4866@nyu.edu

Mei Chen
Center for Data Science
New York University
mc8895@nyu.edu

Jinqian Pan
Center for Data Science
New York University
jp6218@nyu.edu

Raviteja Chukkapalli
Courant Institute of Mathematical Sciences
New York University
rc5124@nyu.edu

Shravan Chaudhari
Courant Institute of Mathematical Sciences
New York University
shravan.c@nyu.edu

Jacopo Cirrone
Center for Data Science
New York University
cirrone@courant.nyu.edu

Abstract

The task of medical image segmentation presents unique challenges, necessitating both localized and holistic semantic understanding to accurately delineate areas of interest, such as critical tissues or aberrant features. This complexity is heightened in medical image segmentation due to the high degree of inter-class similarities, intra-class variations, and possible image obfuscation. The segmentation task further diversifies when considering the study of histopathology slides for autoimmune diseases like dermatomyositis. The analysis of cell inflammation and interaction in these cases has been less studied due to constraints in data acquisition pipelines. Despite the progressive strides in medical science, we lack a comprehensive collection of autoimmune diseases. As autoimmune diseases globally escalate in prevalence and exhibit associations with COVID-19, their study becomes increasingly essential. While there is existing research that integrates artificial intelligence in the analysis of various autoimmune diseases, the exploration of dermatomyositis remains relatively underrepresented. In this paper, we present a deep-learning approach tailored for Medical image segmentation. Our proposed method outperforms the current state-of-the-art techniques by an average of 12.26% for U-Net and 12.04% for U-Net++ across the ResNet family of encoders on the dermatomyositis dataset. Furthermore, we probe the importance of optimizing loss function weights and benchmark our methodology on three challenging medical image segmentation tasks.

1. Introduction

The development of potent CAD (Computer Aided Diagnosis) strategies has been aided by advances in computational power and image analysis algorithms over the past decade. Medical imaging is fundamental to these CAD methods. Obtaining accurate results from CAD techniques relies on acquiring high-quality medical imaging and corresponding annotation. These CAD approaches facilitate various tasks such as image classification, segmentation, spatial mapping, and tracking. Out of these, medical image segmentation is a particularly challenging task due to several complexities. For example, in skin lesion image segmentation, there exists significant intra-class variability and inter-class similarity. This issue is exacerbated by the presence of obscuration and low contrast, which makes the task of separating the affected area from the surrounding image more challenging. On the other hand, sometimes the data required to segment is very complex, with multiple fine-grained and hard-to-segment objects, for example, in the case of histopathology data of dermatomyositis (a kind of autoimmune disease). We provide a few examples of the large variability and low contrast in Figure 1, obscuration in Figure 2, and multiple hard-to-segment small objects in Figure 3.
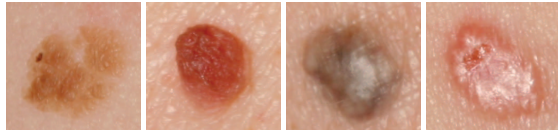
Figure 1. Samples from the dermofit dataset for segmentation, we observe a large lesion color variability from left to right. Interestingly, the background color remains relatively consistent throughout the samples. We also observe a change in contrast from left to right. These confounding factors make the segmentation of skin lesions difficult.
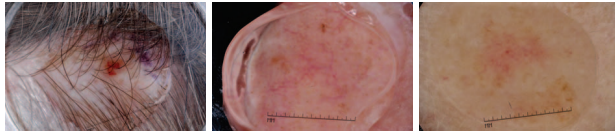

Figure 2. This figure contains samples from the ISIC 2017 dataset. The ISIC-2017 dataset exhibits obscuration and significant intra-class variability, resembling the Dermofit dataset (from Figure 1) regarding inter-class similarity and low contrast (rightmost image).

In addition to these modality-specific complexities, medical imaging datasets are considerably smaller than natural datasets. The main reason for this is the significant expenses and time involved in gathering, annotating medical datasets and privacy concerns. Medical imaging datasets can only be labeled by highly specialized clinicians instead of the possibility of crowdsourced labeling in the case of natural datasets. Privacy concerns pose significant challenges in the open sourcing of medical datasets, particularly for rare or emerging diseases. Medical datasets are typically restricted to institutional use, even when made available [16].

Despite significant progress in medical science, some diseases have not yet been fully comprehended [4]. Autoimmune diseases are a notable category in this context. The lack of a comprehensive catalog of autoimmune diseases, unlike other diseases, is attributed to the diverse nature of their onset and progression [24]. There are still important research questions for autoimmune diseases regarding environmental triggers, pathogenesis, cell inflammation, and interaction. Currently, there are over 80 classified autoimmune diseases. Immune-modulatory drugs are commonly employed for the treatment of autoimmune diseases. However, these drugs have a wide range of effects and lack specificity for autoimmune diseases. Unfortunately, their usage is often linked to other infections and malignant diseases as undesirable side effects. Patients often have limited or no response to these treatments due to the variability within these disorders. So, there is a pressing need for more advanced, fast, and accurate ways to find novel relationships and pathologies that can lead to more effective treatments for autoimmune diseases. To accomplish this, it is imperative to develop precise and adaptable techniques
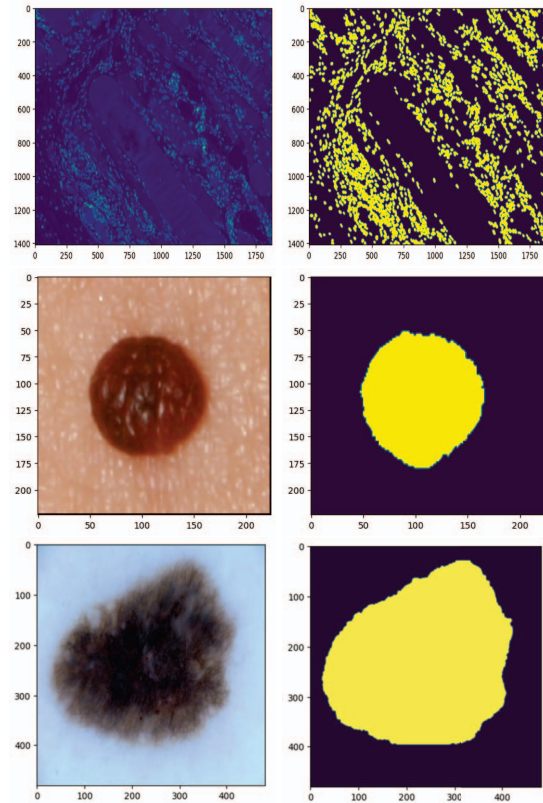

Figure 3. Semantic segmentation task as defined in Section 3 with input image on the left and the corresponding ground truth on the right. On top, we have a sample image (on the left) and the corresponding ground truth (on the right) from the dermatomyositis dataset. Similarly, a sample from the dermofit dataset is in the middle, and a sample from the ISIC-2017 dataset is at the bottom. Unlike a single blob in the skin lesion dataset samples from dermofit and ISIC 2017, we observe that the histopathology whole slide image has many more fine-grained objects with hard-to-segment boundaries. For all the images on the right, the yellow area represents the region of interest (foreground), and the rest is the area other than the region of interest (background).

for analyzing autoimmune diseases related medical images. Implementing AI-based Computer-Aided Diagnosis (CAD) is a potential strategy for achieving this objective. In contrast to other diseases, however, lacking a definitive list and a limited understanding of autoimmune disorders presents a challenge. Consequently, there are few established data collection mechanisms for autoimmune diseases. These factors contribute to the paucity of research on the intersection of autoimmune diseases and CAD approaches. Most extant research in this field is either outdated or lacks open-source methodologies. The study of autoimmune diseases is paramount due to their increasing prevalence[5, 8, 11]. Autoimmune diseases impact a significant portion of the global population, ranging from 5% to 8%. These conditions cause considerable distress to patients and have been found to
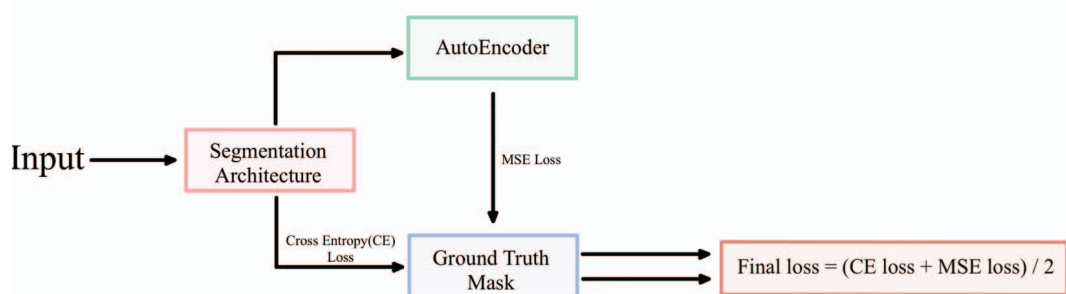
Figure 4. DEDL Architecture[22] with the APP (Autoencoder Post Processing) for dermatomyositis image segmentation as described in Section 2.1. We use this architecture as our baseline and propose changes to it to improve by around 12.26% for U-Net and 12.04% for U-Net++ - this is a considerable improvement over DEDL, as DEDL improved over the previous state-of-the-art approach Van Buren *et al*. [26] by around 5% for segmentation.

have connections with COVID-19, the primary cause of the recent worldwide pandemic [21, 23]. To bridge the divide, Van Buren *et al*. [26] and Singh & Cirrone [22] have made attempts. The main focus of these studies is dermatomyositis. This rare autoimmune disease has received limited attention at the intersection of medical imaging and the application of AI (Artificial Intelligence) for medical image analysis. With this paper,

- We improve upon the existing state-of-the-art approach [22] for dermatomyositis segmentation by an average of 12.26% for U-Net and 12.04% for Unet++ in Section 5.1. Additionally, we benchmark our approach on two other challenging skin-related datasets.

- We study the impact of adding a post-processing autoencoder in addition to U-Net and U-Net++ on three medical imaging datasets in Section 5.1.1.

- We investigate the significance of cross-entropy loss function weights on three challenging medical imaging datasets in Section 5.1.2.

2. Background

Medical image segmentation separates the region of interest, usually a lesion, cells, or other anatomical region of interest, from the slide background. Traditional segmentation processes use pixel-level classification to group pixels into different categories; in the case of semantic segmentation, these would be background and foreground. But with the maturity of Convolutional Neural Networks(CNNs), Ronneberger *et al*. [20] introduced U-Net - an autoencoder-based architecture for biomedical segmentation. The U-Net consists of an encoder and a decoder architecture, where the encoder acts as a feature extractor, and the decoder learns the mask by using the extracted features as input. In addition, the decoder also incorporates the feature maps from the encoder to improve scaling up the representation

to the image mask; these connections are called "Skip-connections." Following U-Net, a wealth of architectures have spawned: U-Net++[27], DeepLab [1], DeepLabV3+ [2], and Feature Pyramid networks (FPN) [12]. All of these architectures build on the autoencoder architecture of U-Net with skip connections. To increase the receptive field of these architectures, various techniques have been introduced, such as dilated networks [18] and nesting architecture, as in the case of U-Net++[27]. Despite these advancements and complex architectures, U-Net remains the choice of architecture for medical image segmentation [17, 22, 26].

2.1. Application of Segmentation Techniques for Autoimmune diseases.

Stafford *et al*. [24] conducted a comprehensive survey to examine the application of AI in the context of autoimmune diseases. They observed the median size of autoimmune datasets is much smaller (99-540 samples per dataset) as compared to datasets pertaining to other medical modalities. The scarce available data poses a significant challenge in acquiring informative priors for artificial intelligence-based CAD approaches on these datasets resulting in sub-par performance. Furthermore, most methodologies for analyzing these datasets are antiquated and lack open-source availability[22]. To overcome these shortcomings Van Buren *et al*. [26] proposed the use of U-Net for segmentation of whole slide images of dermatomyositis histopathology data and open-sourced their approach. Given the considerable size of the whole slide images (WSI) at 1408 × 1876, a tiling approach was employed to partition the WSI into smaller 256 × 256 images, with padding.

They also used a combination of Dice and Binary cross entropy loss to attenuate the problem of pixel distribution imbalance between the area surrounding the region of interest (background pixels) and the region of interest (foreground pixels). Due to this imbalance, the segmentation architecture tends to focus more on the area surrounding the region of interest than the region of interest for segmenta-

Encoder	Technique	U-Net		
		Baseline(w/o APP)	w/ Relu APP	w/ Gelu APP
ResNet-18	DEDL	0.4347	0.4608	0.4788
	Ours	**0.5618**	**0.5479**	**0.5582**
ResNet-34	DEDL	0.4774	0.4467	0.4983
	Ours	**0.5306**	**0.5571**	**0.5606**
ResNet-50	DEDL	0.3798	0.4187	0.3827
	Ours	**0.5556**	**0.5495**	**0.5597**
ResNet-101	DEDL	0.3718	0.4074	0.4402
	Ours	**0.5502**	**0.5678**	**0.5497**

Table 1. Performance comparison of DEDL [22] and our approach on the dermatomyositis dataset for U-Net. We repeat all experiments five times with different seed values and report the IoU on the test in the 95% CI (confidence interval).

Encoder	Technique	U-Net++		
		Baseline(w/o APP)	w/ Relu APP	w/ Gelu APP
ResNet-18	DEDL	0.5274	0.4177	0.4707
	Ours	**0.5622**	**0.5679**	**0.5683**
ResNet-34	DEDL	0.3745	0.4535	0.4678
	Ours	**0.5536**	**0.5685**	**0.5633**
ResNet-50	DEDL	0.4236	0.4685	0.4422
	Ours	**0.5742**	**0.5698**	**0.5514**
ResNet-101	DEDL	0.4311	0.4265	0.4467
	Ours	**0.57**	**0.5727**	**0.5692**

Table 2. Similar to Table 1, in this table, we compare the performance comparison of DEDL [22] and our approach on U-Net++. We report IoU scores averaged over five seed values in the 95% confidence interval (CI) over the test set of the Dermatomyositis dataset in this table.

Dataset	ResNet	U-Net		
		Baseline(w/o APP)	w/ Relu APP	w/ Gelu APP
Dermofit	ResNet18	0.7388	**0.7477**	0.7467
	ResNet34	0.7576	**0.7633**	0.7525
	ResNet50	0.7364	0.7338	**0.7401**
	ResNet101	0.7252	0.7213	**0.7258**
Dermatomyositis	ResNet18	**0.5618**	0.5479	0.5582
	ResNet34	0.5306	0.5571	**0.5606**
	ResNet50	0.5556	0.5495	**0.5597**
	ResNet101	0.5502	**0.5678**	0.5497
ISIC2017	ResNet18	**0.6458**	0.6252	0.6357
	ResNet34	**0.6518**	0.6227	0.6306
	ResNet50	0.605	0.5984	**0.6207**
	ResNet101	0.6267	**0.6325**	0.5884

Table 3. In this table we present the IoU (Intersection over Union) on the test set averaged over five seed values (in 95% CI) for U-Net trained with our proposed technique as mentioned in Section 3.

tion if unattended. Singh & Cirrone [22] further improved on this benchmark by using U-Net and introduced an "Autoencoder Post Processing" (APP) technique. The APP consists of stacked linear layers for the encoder and decoder. This makes the autoencoder much simpler than the convolution and skip connection based U-Net and U-Net++ architectures. After obtaining the mask from a U-Net or U-Net++, it is passed through the APP. Since, the autoencoder consists of only stacked linear layers it creates a noised version of the segmentation output from U-Net and U-Net++.

A mean squared error is then calculated between the autoencoder's output and ground truth. During training, the model trained with the help of the MSE loss (calculated between the autoencoder output and the ground truth) and the cross entropy loss (calculated between the U-Net/U-Net++ output and the ground truth). This helps the model learn a more diverse set of features.

The autoencoder is only used during the training process. Hence there is only a marginal increase in training time while the inference time remains constant. They stud-

Dataset	ResNet	UNet++		
		Baseline(w/o APP)	w/ Relu APP	w/ Gelu APP
Dermofit	ResNet18	**0.744**	0.7408	0.7366
	ResNet34	0.754	0.7553	**0.7599**
	ResNet50	0.737	**0.7408**	0.7379
	ResNet101	0.7232	**0.7264**	0.7229
Dermatomyositis	ResNet18	0.5622	0.5679	**0.5683**
	ResNet34	0.5536	**0.5685**	0.5633
	ResNet50	**0.5742**	0.5698	0.5514
	ResNet101	0.57	**0.5727**	0.5692
ISIC2017	ResNet18	0.6096	**0.6232**	0.6005
	ResNet34	**0.6583**	0.6423	0.6548
	ResNet50	0.6103	**0.6355**	0.619
	ResNet101	0.6018	**0.6164**	0.6041

Table 4. Similar, to table 3 in this table we present IoU averaged over five seed values (in 95% CI) for U-Net++.

Supervision Level	Dataset	Baseline(w/o APP)	w/ Relu APP	w/ GeLu APP
U-Net	Dermofit	0.7395	**0.7415**	0.7413
	Dermatomoyositis	0.5496	**0.5556**	0.5551
	ISIC2017	**0.6323**	0.6197	0.6189
U-Net++	Dermofit	0.7396	**0.7408**	0.7393
	Dermatomoyositis	0.565	**0.5697**	0.5630
	ISIC2017	0.6200	**0.62935**	0.6196

Table 5. Mean IoU in the 95% confidence interval when averaged over the entire ResNet family for U-Net and U-Net++ on the dermofit, dermatomyositis, and the ISIC 2017 dataset. We observe that adding ReLU autoencoder as a post-processing unit improves performance for U-Net++ and, in almost all cases, for U-Net.

ied ReLU and GELU as activation functions for the linear layers of APP and found that ReLU activations work better than GELU. For their choice of architecture, they used U-Net and U-Net++ (nested U-net) with Squeeze and Excitation [10] in the decoder for channel-level attention. To navigate the problem of pixel distribution imbalance between the area surrounding the region of interest (background pixels) and the region of interest (foreground pixels), they used pixel-distribution ratio weights in the cross-entropy loss. Wherein the background pixels used the ratio of background pixels to total pixels as weights, and similarly, the foreground pixels used the ratio of foreground pixels to total pixels as weights. With these changes, they were able to improve on state of the art on the dermatomyositis segmentation task [26] by around 5%. They also suggested a change in evaluation metric from pixel accuracy to IoU (Intersection over Union) as pixel accuracy does not correctly represent the quality of the learned mask as opposed to the ground truth mask. Our study builds upon the foundation established by Singh & Cirrone [22] as a baseline. Our proposed methodology demonstrates significant improvement in performance as compared to the state-of-the-art approach [22]. We achieve an average improvement of 12.26% for U-Net and 12.04% for U-Net++, as elaborated in Section 5.1. Next, we benchmark our methodology on two complex skin lesion datasets in Section 5.1.1.

Furthermore, we investigate the impact of autoencoder for post-processing in Section 5.1.1 and the significance of loss function weights in Section 5.1.2.

3. Methodology

We start with Singh & Cirrone's [22] approach on the dermatomyositis dataset. They use U-Net and U-Net++ as the choice of segmentation architecture with Squeeze and Excitation [10] in the decoder for channel-level attention. Similar to previous studies, our work focuses on semantic segmentation, where the goal is to categorize each pixel in an image into a class. For semantic segmentation, these classes would be - a region of interest (for example, cells in the case of the dermatomyositis dataset) and an area other than the region of interest or background (region other than the cells). For the encoder, we study the performance of the entire ResNet family of CNNs[9]. We use an encoder depth of three, increasing the convolution filter size from 128, 256, to 512. We initialize the encoder with ImageNet pre-trained weights. In the decoder part of U-Net and U-Net++, we use a convolution channel scheme of 256, 128, and 64. For each decoder block, we also use batch normalization as well as squeeze-and-excitation channel excitation after the convolutional layer. As discussed in [22, 26] and Section 2.1, the Dermatomyositis whole slide images contain a lot

more pixels without cells (background) as opposed to with cells (foreground). To attenuate this imbalance in the distribution of pixels, we use **C**ross **D**istribution **W**eights (**CDW**) in the cross-entropy loss. Van Buren *et al.* [26] used random weights. In contrast, Singh & Cirrone [22] used a ratio of the pixel with cells to total pixels and a ratio of non-cell pixels to total pixels as weights for foreground and background, respectively, in the cross-entropy loss. We propose to swap the weights and instead use the ratio of the number of pixels not containing the cell to the total number of pixels as the weight for the foreground. Similarly, the weight of the background is the ratio of the number of pixels containing cells (foreground/object of interest) to the total number of pixels. This alternative weight assignment method aims to enhance the foreground representation. This intuition is very similar to focal loss [13], wherein the misclassification of the minority class is penalized more than that of the minority class. To ensure our results are statistically significant, we conduct all experiments over five different seed values and report the mean values in the 95% confidence interval (C.I) over the five runs. We present these results in Tables 3 and 4. Based on our proposal, we observe an average improvement of 12.26% for U-Net and 12.04% for U-Net++. We further discuss these in Section 5.1. Additionally, we benchmark our approach and study the impact of autoencoder post-processing on two additional challenging dermatology-related datasets - ISIC 2017 and the dermofit dataset in Section 5.1.1. Both datasets are challenging due to large intra-class variations and inter-class similarities as depicted in Figure 1 and obscuration in 2. Finally, we study the impact of using mean-frequency weights and compare the results with distribution-swapped weights for U-Net and U-Net++ over the ResNet family of encoders and three datasets in Section 5.1.2.

4. Experimental Details

4.1. Datasets

We use a 70-10-20 split for the dermatomyositis and dermofit datasets. For the ISIC-2017 dataset, we use the same splits as used in the 2017 ISIC competition.

Dermatomyositis: We use the same dataset as used in previous works on dermatomyositis segmentation [22, 26]. To give an idea about the modality of the dataset, we show a random sample from the test set in Figure 3. The Dermatomyositis dataset is collected from 198 muscle biopsies collected from seven dermatomyositis patients. These files are then stored in TIFF format. Each TIFF image contains eight slides that indicate the presence or absence of phenotypic markers by setting binary thresholds for each channel (1-DAPI, 2-CXCR3, 3-CD19, 4-CXCR5, 5-PD1, 6-CD4, 7-CD27, 8-Autofluorescence). For segmentation, we used the

DAPI-stained image. Each whole slide image was tiled into 480x480. We further expand on this in Section 4.1.2. This is a particularly challenging dataset due to the large number of fine-grained objects (cells) to be segmented per image, as discussed in Section 1.

Dermofit [7]: As shown in Figure 3, the Dermofit dataset contains 1300 skin lesion RGB images. These data are taken with a high-quality SLR camera in controlled (ring flash) indoor illumination. The Dermofit dataset contains ten categories; each includes a different number of instances: Actinic Keratosis (AK): 45, Basal Cell Carcinoma (BCC): 239, Melanocytic Nevus / Mole (ML): 331, Squamous Cell Carcinoma (SCC) sample 88, Seborrhoeic Keratosis (SK): 257, Intraepithelial carcinoma (IEC): 78, Pyogenic Granuloma (PYO): 24, Haemangioma (VASC): 96, Dermatofibroma (DF): 65, Melanoma (MEL): 76. No two images in this dataset are of the same size, as a preprocessing step we interpolate all images to 480x480 and then resize to 224x224 to ensure uniformity.

ISIC Challenge 2017 Dataset, Lesion Segmentation Task [3]: The International Skin Imaging Collaboration (ISIC) is a large publicly accessible dataset. We show a sample from the test set in Figure 3. In our case, we use the segmentation dataset from 2017 and use the original splits wherein 2,000 images were used as training, 150 images as validation, and 600 images as the test set. The ISIC 2017 and the Dermofit datasets described above are skin lesion datasets with high intra-class variability and inter-class similarities with obscuration areas of interest, as discussed in Section 1.

4.1.1 Common Implementation Details

We implemented all models in Pytorch [19] using a single NVIDIA RTX-8000 GPU with 64 GB RAM and 3 CPU cores. All models are trained with an Adam optimizer with an initial learning rate (lr) of 3.6e-4 and a weight decay 1e-5. We use a cosine annealing scheduler with a maximum of 50 iterations and a minimum learning rate of 3.4e-4 to adjust the learning rate based on each epoch. We train all architectures for 50 epochs with batch size 16, followed by testing on a held-out set. We use IoU (Intersection over Union) as our evaluation metric on the test set. This aligns with previous work by Singh & Cirrone [22]. We repeat all experiments with five different seed values and report the mean value in the 95% confidence interval in all tables.

4.1.2 Data-Preprocessing

Images of the dermatomyositis dataset have a uniform size of 1408×1876; we tiled each image into 12 sub-images of size 480×480 inline with previous work [22]. In contrast,

Encoder	Technique	U-Net			U-Net++		
		Baseline*	w/ ReLU APP	w/ GELU APP	Baseline*	w/ ReLU APP	w/ GELU APP
ResNet-18	CDW	0.5618	0.5479	**0.5582**	0.5622	**0.5679**	0.5683
	Mean Frequency	**0.5645**	**0.5592**	0.5405	**0.5852**	0.5603	**0.5814**
ResNet-34	CDW	0.5306	**0.5571**	0.5606	0.5536	0.5685	0.5633
	Mean Frequency	**0.5555**	0.5551	**0.5616**	**0.5729**	**0.5763**	**0.57**
ResNet-50	CDW	**0.5556**	0.5495	**0.5597**	**0.5742**	0.5698	0.5514
	Mean Frequency	0.5512	**0.5652**	0.5585	0.57	**0.5723**	**0.5929**
ResNet-101	CDW	0.5502	**0.5678**	0.5497	0.57	**0.5727**	0.5692
	Mean Frequency	**0.5506**	0.5537	**0.5596**	**0.5892**	0.5678	**0.5773**

Table 6. This table displays the average Intersection over Union (IoU) values obtained from five separate runs on the dermatomyositis test set, with their corresponding 95% confidence intervals (CI). In this context, CDW refers to the utilization of cross-distribution weights in the calculation of cross-entropy loss. Here, Baseline* represents the use of the segmentation architecture with autoencoder for post-processing (APP).

Encoder	Technique	U-Net			U-Net++		
		Baseline*	w/ ReLU APP	w/ GELU APP	Baseline*	w/ ReLU APP	w/ GELU APP
ResNet-18	CDW	0.7388	0.7477	**0.7467**	0.744	0.7408	0.7366
	Mean Frequency	**0.75**	**0.7498**	0.7377	**0.7469**	**0.7413**	**0.7449**
ResNet-34	CDW	**0.7576**	**0.7633**	0.7525	0.754	0.7553	**0.7599**
	Mean Frequency	0.7533	**0.7633**	**0.7535**	**0.7602**	**0.7635**	0.7547
ResNet-50	CDW	0.7364	**0.7338**	**0.7401**	**0.737**	**0.7408**	0.7379
	Mean Frequency	**0.7379**	0.731	0.7385	0.7362	0.7358	**0.7411**
ResNet-101	CDW	**0.7252**	0.7213	0.7258	**0.7232**	**0.7264**	0.7229
	Mean Frequency	0.7212	**0.7242**	**0.7236**	0.7156	0.7247	**0.7234**

Table 7. Similar to Table 6, in this table, we present the IoU averaged over five runs in 95% confidence interval on the dermofit test set for U-Net and U-Net++. Like Table 6, CDW represents the scenario when cross-distribution weights are used for the cross-entropy loss and Baseline* represents the use of the segmentation architecture with autoencoder for post-processing (APP).

the Dermofit and the ISIC2017 datasets contain images of different sizes, i.e., no two images in the dataset are the same size. Additionally, since the other two datasets (dermofit and ISIC-2017) contain skin lesions, they have significantly denser and larger mask labels than the dermatomyositis dataset. Thus, a different image preprocessing step is applied to the latter two datasets: bilinear interpolation to 480×480 followed by a resize to 224×224. For augmentation, we use the same set of augmentation as used in Singh & Cirrone's work [22], along with Red channel normalization or "Rnorm" [25] for all of our experiments.

5. Results and Discussion

5.1. Improvement over the current state-of-the-art for Dermatomyositis WSI Segmentation [22]

Following the methodology (Section 3) and experimentation setup (Section 4), we present the IoU averaged over five runs in the 95% confidence interval on the test set for the Dermatomyositis dataset in Table 1 for U-Net and in Table 2 for U-Net++. We observe that our approach improves over Singh & Cirrone's approach (DEDL) [22] consistently over the entire ResNet family for baseline as well as with APP (both ReLU and GELU based) for both U-Net and U-Net++. When averaged over the ResNet family of encoders and the three paradigms (baseline approach without using autoencoders, ReLU autoencoders, and GELU-based autoencoders), we observed that our approach improves over the previous state-of-art [22] for Dermatomyositis segmentation by 12.26% and 12.04% for U-Net and U-Net++ respectively.

5.1.1 Impact of Incorporating Autoencoder Post-Processing.

As described in Section 2.1, Singh and Cirrone [22] introduced an "Autoencoder Post Processing" unit or APP after the main segmentation architecture. The purpose of this autoencoder was to provide a more noised version of the prediction from the U-Net or U-Net++. The mean square error loss between the noised output and the ground truth mask, along with the weighted-cross entropy loss between the output of the U-Net or U-Net++ and the ground truth, is optimized during training. This is depicted in Figure 4. They studied the impact of using APP with ReLU and GELU activations only on the Dermatomyositis dataset. In this section, with our improved approach as presented in Section 3, we study the impact of adding APP on two ad-

Encoder	Technique	U-Net			U-Net++		
		Baseline*	w/ ReLU APP	w/ GELU APP	Baseline*	w/ ReLU APP	w/ GELU APP
ResNet-18	CDW	**0.6458**	0.6252	**0.6357**	0.6096	**0.6232**	0.6005
	Mean Frequency	0.6257	**0.6394**	0.6307	**0.6177**	0.6198	**0.6074**
ResNet-34	CDW	**0.6518**	0.6227	**0.6306**	**0.6583**	0.6423	**0.6548**
	Mean Frequency	0.6314	**0.6409**	0.6322	0.6412	**0.6454**	0.6513
ResNet-50	CDW	0.605	0.5984	0.6207	0.6103	**0.6355**	0.619
	Mean Frequency	**0.6396**	**0.6223**	**0.6337**	**0.6354**	0.628	**0.646**
ResNet-101	CDW	**0.6267**	**0.6325**	0.5884	0.6018	**0.6164**	**0.6041**
	Mean Frequency	0.6137	0.6283	**0.6175**	**0.6049**	0.6112	0.6016

Table 8. In this table, we showcase the IoU results for U-Net and U-Net++ on the ISIC 2017 test set. The values represent the average IoU over five runs in the 95% confidence interval. CDW, in this context, refers to the utilization of cross-distribution weights for cross-entropy loss, as demonstrated in Tables 6 and 7. Baseline* represents the use of the segmentation architecture with autoencoder for post-processing.

Dataset	CDW	Median Frequency
Dermofit	0.7395	**0.7406**
DM*	0.5496	**0.5564**
ISIC2017	0.6197	**0.6276**

Table 9. In this table, we present the average IoU (in the 95% confidence interval) over the ResNet family of encoders and the three paradigms (Baseline, with ReLU and GELU APP) for U-Net over the dermatomyosytis, the dermofit, and ISIC 2017 test sets from Tables 6, 7 and 8 respectively. Here, DM* represents the Dermatomyosytis dataset, and CDW represents Cross Distribution Weight.

Dataset	CDW	Median Frequency
Dermofit	0.7396	**0.7397**
DM*	0.565	**0.5793**
ISIC2017	0.62	**0.6248**

Table 10. Similar to Table 9, In this table, we provide the average Intersection over Union (IoU) values (in 95% CI) for the ResNet family for U-Net++ architecture on three different test sets: Dermatomyosytis, Dermofit, and ISIC 2017.

Dataset	CDW	Median Frequency
Dermofit	[0.3037, 0.6963]	[0.7180, 1.6466]
DM*	[0.1479, 0.8521]	[0.5986, 3.0348]
ISIC2017	[0.2020, 0.7980]	[0.6265, 2.4755]

Table 11. Cross-entropy weights used for experimentation as described in Section 5.1.1. Here, DM* represents the Dermatomyosytis dataset as described in Section 4.1 and CDW represents Cross Distribution Weight.

ditional challenging dermatology datasets. We present the IoU over the test set in the 95% confidence interval over the ResNet family of encoders in Tables 3 and 4 for ISIC 2017 and the Dermofit dataset, respectively. Adding ReLu and GeLU-based APP improves performance over the baseline architecture (with no APP) for U-Net and U-Net++ in most cases for the Dermofit. To better understand the result, we average the IoU on the test set over the entire ResNet family for U-Net and U-Net++ and present the results in Table 5. From Table 5, we observe that the addition of APP, es-

pecially ReLU-based APP, does improve performance over the baseline (not using APP) in almost all cases for U-Net++ and U-Net. The addition of APP is did not improve performance only in the case of the ISIC-2017 dataset for U-Net.

5.1.2 Impact of Cross-entropy loss weights

In section 3, we explained our rationale for switching from distribution-based weights to cross-distribution-based weights for the cross-entropy loss. In this section, we study the impact of changing the cross-entropy weights from cross-distribution-based weights to mean frequency weights[6]. The median frequency weight received by each class is derived from the reciprocal of the pixel ratio of a particular class, normalized by the median frequency of the class[14, 15]. The median frequency and the cross-distribution weights, calculated over our three datasets, are mentioned in Table 3. Mathematically, median frequency weights (w_c) are defined as follows: $w_c = \frac{\text{med\_freq}}{n_c}$. Here, n_c is the number of pixels belonging to class c in the training dataset, and med_freq is the median of the frequency of pixels belonging to each class in the dataset.[1] Where $n_c = \frac{\alpha}{\beta}$, here, α represents the number of pixels of a class, and β represents the total number of pixels in images where the given class is present. We compare the weights calculated by cross-distribution and median frequency in Table 3. We provide the full comparative result of using cross-distribution weights and mean frequency over the three datasets in the 95% confidence interval averaged over five seed values Tables 6, 7 and 8 for the Dermatomyosytis, Dermofit, and the ISIC-2017 datasets, respectively. Additionally, to summarize these results, we present the average over the ResNet family and the three training paradigms (baseline without APP and APP with ReLU and GELU layers) in Tables 9 and 10 for U-Net and U-Net++, respectively. From these tables, we observe that median-frequency weights for cross-entropy loss improve performance over cross-distribution

---

[1]In our case, there are only two classes - foreground (region of interest) and background (area other than the region of interest).

weights, although the improvement is marginal in almost all cases.

6. Conclusion

We observed that our approach of using Cross Distribution Weights (CDW) improved segmentation performance over the previous state-of-the-art approach for dermatomyositis segmentation [22] by 12.26% for U-Net and by 12.04% for U-Net++ averaged over the ResNet family. Furthermore, adding APP (Autoencoder Post Processing) improves segmentation performance marginally in the case of dermatomyositis and dermofit datasets. In the case of the ISIC 2017 dataset, the addition of APP is only useful in the case of U-Net++. We have open-sourced our approach at `https://github.com/pranavsinghps1/Enhancing-Medical-Image-Segmentation`. We hope that our study and open-sourced approach will catalyze further research at the intersection of autoimmune diseases like dermatomyositis and the application of AI as well as for other dermatology-related datasets. This would help us better understand the immunology of autoimmune diseases and answer some of the critical research questions to develop improved healthcare solutions.[2]

---

[2]Potential negative societal impact: Autoimmune diseases are extremely heterogeneous; the dermatomyositis dataset used in our experiments is geographically restricted. Hence, this is a study of a particular variant. This study might or not be generalizable for other variants. Hence application on a wider scale for real-life scenarios should only be trusted after clearance from the concerned health and safety governing bodies.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan 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):834–848, 2017.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[3] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[4] Glinda S Cooper, Milele LK Bynum, and Emily C Somers. Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. *Journal of autoimmunity*, 33(3-4):197–207, 2009.

[5] Michael Ehrenfeld, Angela Tincani, Laura Andreoli, Marco Cattalini, Assaf Greenbaum, Darja Kanduc, Jaume Alijotas-Reig, Vsevolod Zinserling, Natalia Semenova, Howard Amital, et al. Covid-19 and autoimmunity. *Autoimmunity reviews*, 19(8):102597, 2020.

[6] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *CoRR*, abs/1411.4734, 2014.

[7] Robert Fisher and Jonathan Rees. Dermofit project datasets. 2017.

[8] Caroline Galeotti and Jagadeesh Bayry. Autoimmune and inflammatory diseases following covid-19. *Nature Reviews Rheumatology*, 16(8):413–414, 2020.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[11] Aaron Lerner, Patricia Jeremias, and Torsten Matthias. The world incidence and prevalence of autoimmune diseases is increasing. *Int J Celiac Dis*, 3(4):151–5, 2015.

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[14] Yahui Liu, Jian Yao, Xiaohu Lu, Menghan Xia, Xingbo Wang, and Yuan Liu. Roadnet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):2043–2056, 2019.

[15] Yahui Liu, Jian Yao, Xiaohu Lu, Renping Xie, and Li Li. Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing*, 338:139–153, 2019.

[16] Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith. What makes transfer learning work for medical images: feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9225–9234, 2022.

[17] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann Elmore, and Linda Shapiro. Y-net: Joint segmentation and classification for diagnosis of breast biopsy images. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2018.

[18] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[21] Chetan Sharma and Jagadeesh Bayry. High risk of autoimmune diseases after covid-19. *Nature Reviews Rheumatology*, pages 1–2, 2023.

[22] Pranav Singh and Jacopo Cirrone. A data-efficient deep learning framework for segmentation and classification of histopathology images. In *European Conference on Computer Vision*, pages 385–405. Springer, 2022.

[23] Anuroop Sriram, Matthew Muckley, Koustuv Sinha, Farah Shamout, Joelle Pineau, Krzysztof J Geras, Lea Azour, Yindalon Aphinyanaphongs, Nafissa Yakubova, and William Moore. Covid-19 prognosis via self-supervised representation learning and multi-image prediction. *arXiv preprint arXiv:2101.04909*, 2021.

[24] IS Stafford, M Kellermann, E Mossotto, Robert Mark Beattie, Ben D MacArthur, and Sarah Ennis. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ digital medicine*, 3(1):30, 2020.

[25] Dang N.H. Thanh, Uğur Erkan, V.B. Surya Prasath, Vivek Kumar, and Nguyen Ngoc Hien. A skin lesion segmentation method for dermoscopic images based on adaptive thresholding with normalization of color models. In *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE)*, pages 116–120, 2019.

[26] Kayla Van Buren, Yi Li, Fanghao Zhong, Yuan Ding, Amrutesh Puranik, Cynthia A Loomis, Narges Razavian, and Timothy B Niewold. Artificial intelligence and deep learning to map immune cell types in inflamed human tissue. *Journal of Immunological Methods*, 505:113233, 2022.

[27] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.