cancers

MDPI

Article

Contrastive Clustering-Based Patient Normalization to Improve Automated In Vivo Oral Cancer Diagnosis from Multispectral Autofluorescence Lifetime Images

Kayla Caughlin [1], Elvis Duran-Sierra [2], Shuna Cheng [2], Rodrigo Cuenca [3], Beena Ahmed [4], Jim Ji [5], Mathias Martinez [6], Moustafa Al-Khalil [6], Hussain Al-Enazi [7], Javier A. Jo [3] and Carlos Busso [1,8,*]

[1] Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX 75080, USA; krc170830@utdallas.edu
[2] Department of Biomedical Engineering, Texas A&M University, College Station, TX 77840, USA; eduran3@mdanderson.org (E.D.-S.); csncbmp@gmail.com (S.C.)
[3] School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK 73019, USA; rodrigo.cuenca@ou.edu (R.C.); javierjo@ou.edu (J.A.J.)
[4] School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2033, Australia; beena.ahmed@unsw.edu.au
[5] Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha 23874, Qatar; jim.ji@qatar.tamu.edu
[6] Department of Cranio-Maxillofacial Surgery, Hamad Medical Corporation, Doha 3050, Qatar; mathias_martinezcoronel@urmc.rochester.edu (M.M.); malkhalil@hamad.qa (M.A.-K.)
[7] Department of Otorhinolaryngology Head and Neck Surgery, Hamad Medical Corporation, Doha 3050, Qatar; halenazi@hamad.qa
[8] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
* Correspondence: busso@cmu.edu

Simple Summary: Lip and oral cavity cancer caused over 177,000 deaths globally in 2020, but patient survival increases with earlier diagnosis. One barrier to early diagnosis is the invasive nature of biopsies needed for diagnosis. Automated diagnosis systems have the potential to perform non-invasive diagnosis by pairing novel imaging data with deep learning models. Given the variability between patients, access to a sufficiently large training database from human subjects limits deep learning applications. We propose a model that maps non-invasive images of oral tissue to a diagnosis by encouraging the model to group normal samples close together (reducing variability between patients). Our model improves non-invasive oral cancer diagnosis through a robust training process that only requires a small amount of data. This work shows how we can address small data challenges through model architecture and training, rather than through the collection of larger databases or manual corrections and normalizations.

Abstract: Background: Multispectral autofluorescence lifetime imaging systems have recently been developed to quickly and non-invasively assess tissue properties for applications in oral cancer diagnosis. As a non-traditional imaging modality, the autofluorescence signal collected from the system cannot be directly visually assessed by a clinician and a model is needed to generate a diagnosis for each image. However, training a deep learning model from scratch on small multispectral autofluorescence datasets can fail due to inter-patient variability, poor initialization, and overfitting. **Methods:** We propose a contrastive-based pre-training approach that teaches the network to perform patient normalization without requiring a direct comparison to a reference sample. We then use the contrastive pre-trained encoder as a favorable initialization for classification. To train the classifiers, we efficiently use available data and reduce overfitting through a multitask framework with margin delineation and cancer diagnosis tasks. We evaluate the model over 67 patients using 10-fold cross-validation and evaluate significance using paired, one-tailed *t*-tests. **Results:** The proposed approach achieves a sensitivity of 82.08% and specificity of 75.92% on the cancer diagnosis task with a sensitivity of 91.83% and specificity of 79.31% for margin delineation as an auxiliary task. In comparison to existing approaches, our method significantly outperforms a *support vector machine* (SVM) implemented with either *sequential feature selection* (SFS) (p = 0.0261) or L1 loss (p = 0.0452)

check for updates

Citation: Caughlin, K.; Duran-Sierra, E.; Cheng, S.; Cuenca, R.; Ahmed, B.; Ji, J.; Martinez, M.; Al-Khalil, M.; Al-Enazi, H.; Jo, J.A.; et al. Contrastive Clustering-Based Patient Normalization to Improve Automated In Vivo Oral Cancer Diagnosis from Multispectral Autofluorescence Lifetime Images. *Cancers* **2024**, *16*, 4120. https://doi.org/10.3390/cancers16234120

Academic Editor: René-Jean Bensadoun

Received: 22 October 2024
Revised: 20 November 2024
Accepted: 3 December 2024
Published: 9 December 2024

when considering the average of sensitivity and specificity. Specifically, the proposed approach increases performance by 2.75% compared to the L1 model and 4.87% compared to the SFS model. In addition, there is a significant increase in specificity of 8.34% compared to the baseline autoencoder model (p = 0.0070). **Conclusions:** Our method effectively trains deep learning models for small data applications when existing, large pre-trained models are not suitable for fine-tuning. While we designed the network for a specific imaging modality, we report the development process so that the insights gained can be applied to address similar challenges in other non-traditional imaging modalities. A key contribution of this paper is a neural network framework for multi-spectral fluorescence lifetime-based tissue discrimination that performs patient normalization without requiring a reference (healthy) sample from each patient at test time.

Keywords: multispectral autofluorescence lifetime imaging; automated cancer diagnosis; margin delineation; patient normalization; regularization; deep learning

1. Introduction

In 2020, lip and oral cancer caused over 177,000 deaths globally [1]. In particular, countries with lower *human development indices* (HDIs) saw increased incidence and mortality rates for lip and oral cancer [1]. Efforts to develop automated oral cancer classifiers result from the clinical need to easily and reliably classify oral lesions, especially in low HDI regions. Both margin delineation (classification of lesion tissue versus healthy tissue) and diagnosis (classification of lesion tissue as benign or malignant) have clinical relevance. While diagnosis classifiers are useful for determining if treatment or resection is required, margin delineation classifiers are useful in image-guided surgery to ensure the margin of the lesion is removed during resection. We are interested in *multispectral fluorescence lifetime imaging* (maFLIM) since this imaging modality is fast and non-invasive. In contrast to whole-slide images and mobile camera images (RGB format), we note that maFLIM images in our domain provide in vivo metabolic and biochemical information in the form of three biexponential-like decays per pixel. As a result, commonly used medical image processing architectures such as U-Net [2] and networks pre-trained on ImageNet (e.g., [3–5]) cannot simply be fine-tuned for our application. Similarly, typical data augmentation methods such as scaling, flipping, or rotation do not apply to our domain.

A key challenge in developing an automated diagnosis or delineation classifier in maFLIM is the variability of oral lesions. Within lesions of the same class, tissue characteristics from different patients may differ due to the lesion location (e.g., tongue versus tonsil) [6]. In addition, oral cancers are associated with a variety of different risk factors, including *human papillomavirus* (HPV) and tobacco use [7]. To reduce inter-patient variability, some authors have proposed incorporating a reference (healthy) and suspicious (potentially malignant) example for each patient [8,9]. Adversarial methods have also been used to reduce inter-patient variability [10]. One patient normalization approach in our domain calculates features for the healthy and suspicious images and uses the difference of each feature's healthy and suspicious values as the input to a classifier [8,9]. Outside of a reference value for patient normalization, the healthy samples are not typically used when training networks for oral cancer diagnosis [8,9]. Similarly, margin delineation approaches may only compare malignant and healthy samples, discarding the benign images [11,12].

In this work, we explore two main questions: (1) Can we achieve a reduction in inter-patient variability without direct correspondence between healthy and suspicious images from each patient? (2) Can we combine the two clinically relevant tasks of margin delineation and cancer diagnosis in a beneficial way? Our method consists of two steps: contrastive pre-training followed by the addition of two task classifiers. Samples from three classes are used: benign, malignant, and healthy. The contrastive pre-training step trains an encoder that embeds each pixel such that pixels from the same class cluster together, away from pixels from a different class (see bottleneck representation in Figure 1). Through

the contrastive pre-training, we force the encoder to focus on characteristics specific to the class, rather than specific to a particular patient. Instead of normalizing each lesion feature by subtracting a paired healthy feature value, our approach uses contrastive learning to create patient-invariant representations of each image, whether healthy or suspicious. Through our patient normalization, we reduce inter-patient variability without any direct pairing between a patient's healthy and suspicious images. In addition to the contrastive pre-training, we add task classifiers for the margin delineation and diagnosis tasks. The multitask framework provides enough regularization to the network to avoid the use of an autoencoder as reported in Caughlin et al. [13] and allows both clinically relevant tasks to be accomplished with the same encoder.

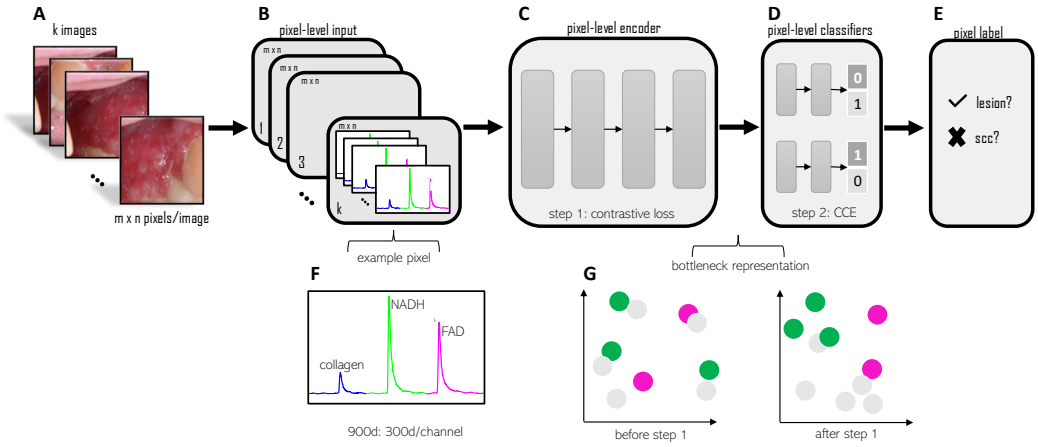


Figure 1. Approach overview. Step 1 (**A**–**C**)—input pixel-level training data (see example pixel panels (**B**,**F**)) into the encoder and train using contrastive loss. After step 1 pre-training, the bottleneck representation shows the clustering of each class (see bottleneck representation Panel (**G**)). Step 2 (**D**,**E**)—add pixel-level classifiers and train with categorical cross-entropy losses. Aggregate all pixel-level labels from a single image using a 50% threshold to label each image.

Our approach improves diagnosis performance in comparison to competitive baselines. In addition, our network achieves good performance on the margin delineation task, even though delineation is treated as an auxiliary task within our framework. Our results show that training a multitask model can help regularize the training process for several clinically relevant tasks, while contrastive learning can reduce inter-patient variability (a key challenge in automated medical tasks). Our full cancer diagnosis model achieves a sensitivity of 82.08% and a specificity of 75.92%. The improvements over baselines range from 1.46% to 4.87% considering the average of sensitivity and specificity. In summary, the primary contributions of our work are implicit patient normalization using a contrastive pre-training step and further regularization through a novel multitask network for margin delineation and cancer diagnosis of oral lesions.

The rest of the paper is organized as follows: Section 2 describes related work in terms of maFLIM classification, contrastive learning, and multitask learning. Section 3 details the loss functions and training process. Section 4 analyzes the contrastive loss function and model performance, including baseline comparison and model analysis. Section 5 summarizes the results and suggests further research directions.

2. Related Work

2.1. Classification Approaches in maFLIM

Tasks for automated classification of oral lesions using maFLIM include margin delineation and cancer diagnosis. In margin delineation, the task is to distinguish between healthy tissue versus cancerous tissue. In cancer diagnosis, the task is to determine if a lesion is benign or cancerous. Each task is typically treated separately using a machine

learning approach, with variations on lifetime and intensity features input into each classifier. For example, Duran et al. [11] evaluated *support vector machines* (SVMs), *quadratic discriminant analysis* (QDA), and ensemble approaches on the margin delineation task. They used only healthy, cancerous, and pre-cancerous training data, without incorporating any images from benign lesions, and did not consider the cancer diagnosis task. Jo et al. [8] used similar classifiers for the cancer diagnosis task, where the classifiers were trained on patient-normalized features from benign, cancerous, and pre-cancerous lesions. While healthy samples were not included as a separate class, the patient normalization step required each patient in both the training set and the testing set to have two images: one of the lesion and one of the healthy tissue [8]. Similarly in the skin cancer domain, Vasanthakumari et al. [9] used SVM, QDA, and *linear discriminant analysis* (LDA) classifiers trained with phasor features for cancer diagnosis of skin lesions.

Transitioning to deep learning approaches, Marsden et al. [6] implemented a pre-trained *convolutional neural network* (CNN) for the margin delineation task. In cancer diagnosis, we previously introduced an autoencoder and classifier approach [13]. However, our previous work only considered the benign and cancerous or pre-cancerous samples during training and applied no patient normalization. While our approach improved classification performance over machine learning baselines, the addition of patient normalization could increase classification performance. While patient normalization could be accomplished in a variety of ways, we focus on a contrastive pre-training step without direct correspondence between healthy and lesion images from the same patient.

2.2. Contrastive Learning Background

The primary objective of typical contrastive losses is to map the input to a space where the distances between samples are meaningful for a task [14–17]. The variations consist of differences in the number of positive and negative examples used (i.e., pairs, triplets, or N-pairs) or in the way that samples are grouped (i.e., from class labels, data augmentations, or temporal separation).

Considering the number of positives and negatives, Chopra et al. [18] described a loss function based on a Siamese network, where a sample was contrasted with either a positive or a negative example using the L1 distance. The model learned to map images of the same person into a space where the images are close together. Images of different people were mapped to a space where the distance between the images was large [18]. In addition to the positive pair in Chopra's contrastive loss, the triplet loss also considers a negative example for each input [14,15]. The triplets consist of a sample (the anchor), an example from the same class as the anchor (the positive), and an example from a different class (the negative). Weinberger et al. [14] and Schroff et al. [15] used triplets of samples in a hinge-based loss. In the hinge loss approach, the model is penalized when the distance between two samples with different labels does not exceed the distance between two samples of the same label by a margin [14,16]. In Weinberger et al. [14], the triplets were formed using the k neighbors, and an additional term was added to cluster the anchor and positive example.

Following contrastive and triplet losses, Sohn [16] introduced the N-pair loss. The N-pair loss modifies the triplet loss to consider *N-1* negative samples for each anchor [16]. By using more negative examples, the loss is less dependent on hard negative mining, shows faster convergence, and leads to improved performance [16]. Moving further, Khosla et al. [17] introduced a loss function for contrastive pre-training using class labels. Their formulation uses the class labels of the samples in each batch to form positives and negatives, where each positive example in the batch and each negative example in the batch contributes to a sample's loss. Using a two-stage training process, the results from Khosla et al. [17] showed that a contrastive pre-training step improves classification performance over a single cross-entropy classifier training stage, even though both steps are fully supervised.

In the medical imaging domain, various losses and pairs strategies have been explored [10,19–21]. Choudary et al. [20] used a triplet loss for *image quality assessment*

(IQA) on whole slide images, starting with a pre-trained CNN. Contrastive learning using different images from the same patient has been used for X-ray images [21]. Similarly, augmentations of the same signal have been used for contrastive learning in *electroencephalogram* (EEG) and *electrocardiogram* (ECG) tasks [10]. Both of these frameworks used the InfoNCE loss [10,21]. Goswami et al. [19] used losses similar to ours to reduce inter-patient variability in leukemia classification. The authors trained on a dataset of images from microscopes, enabling the use of a standard CNN architecture pre-trained on ImageNet [19]. We note that raw maFLIM data are not compatible with pre-trained CNNs from natural images, so novel approaches are needed, especially with reduced training data.

2.3. Multitask Learning Background

Rather than training on a single task, multitask learning uses a shared hidden layer that connects to more than one task. Caruana [22] showed that training with more than one task simultaneously improves generalization and allows for the learning of different features. The specific formulation of these extra tasks varies; they could involve predicting features present in the training set but absent in the test set [22]. The extra tasks could also involve predicting future values [22]. In these formulations, the extra tasks are used solely to guide training [22]. Alternatively, all the training tasks could be utilized during both training and testing phases [22].

In our case, we used two related tasks: margin delineation and cancer diagnosis. In general, we found that the margin delineation (distinction between malignant versus healthy) is easier than the classification task (distinction between malignant versus benign). In a setting where some tasks are easier than others, Caruana hypothesized that the easier tasks can help identify relevant features that the more difficult task might struggle to learn [22]. The multitask framework can also learn features that neither task alone would identify [22]. Finally, multitask learning also uses small datasets efficiently because more information is available during training on multiple tasks (called "data amplification" by Caruana [22]).

In the bioimaging domain, various multitask formulations have led to improvements in automated cancer diagnosis [23–25]. Seo et al. [23] used multitask learning to automatically locate tumors in lung or liver *computed tomography* (CT) images. The authors train a CNN using as few as 48 images. They attributed the success of their method, in part, to the increased regularization of the network from multitask training [23]. Similarly, Khosravan and Bagci [24] trained a CNN on CT scans to find lung nodules and to determine if a proposed region is a lung nodule ("false positive reduction"). The authors emphasize the benefits of multitask learning for improved generalization, discriminative feature identification, and training on small data [24]. Sainz et al. [25] used a multitask network to diagnose breast cancer and to locate lesions and calcifications from mammograms.

While the same multitask learning insights apply to our research, we note that our imaging modality is completely different from CT or mammography. Common imaging modalities like CT and mammography show anatomical information and are typically processed by the network on an image level. In contrast, maFLIM data primarily shows functional information related to the tumor microenvironment. Furthermore, we input pixel-level information into our network. As shown in Figure 1 Panel F, each pixel contains a time-series sequence.

3. Materials and Methods

3.1. Motivation

Existing patient normalization approaches directly compare pairs of images from each patient: a healthy reference sample and a suspicious/lesion sample. In contrast, we want to incorporate patient normalization into the training process. The network will learn to remove patient-specific characteristics and identify common class-specific features across patients, without paired comparison of each patient to their reference sample. Ideally, the contrastive pre-training step organizes the feature space such that each class groups together

far from the other classes (see bottleneck representation, Figure 1). Building on the success of multitask learning, we want to add classifiers that reduce overfitting and allow the primary and auxiliary tasks to benefit from each other, leveraging a favorable initialization from the pre-training step. We hypothesize that the contrastive plus multitask framework will reduce inter-patient variability, smooth the decision boundaries, and regularize the network to improve generalization *without requiring two samples from each patient at test time*.

3.2. Contrastive Pre-Training

As stated in a recent survey, "contrastive learning is an idea, not a specific model [26]". While contrastive learning is often used to generate the representation of entire images, the concept is not exclusive to images. It has been used in other modalities as well, such as text summarization [27], graph data [28], and video [29]. We note that in contrastive learning with images, the loss is generally applied to the embedding (hidden layer) rather than directly to the image input [26]. Similarly, our encoder maps the 900-dimensional pixel input (containing three biexponential decays in a single pixel) to a lower-dimensional embedding where we can apply a contrastive loss. The contrastive loss reduces variability by bringing the reference (normal) pixels closer together in the embedding space. We can think of this process as a learned normalization. Fernando et al. [30] discussed how natural variability and measurement variability in patient data lead to challenges in medical anomaly detection, providing an example of patient endoscopy images that belong to separate classes but are visually similar. Similarly, in the presence of high inter-patient variability, the distance between an individual patient's normal and lesion data may be smaller than the distance between two different patients' normal images. In this setting, the model may reach a local optimum by overfitting to individual data points rather than learning generalizable characteristics. Using the contrastive loss, we guide the model to ignore patient-specific characteristics that are irrelevant to the task of cancer diagnosis. This contrastive pre-training creates a better decision boundary, where the model learns to focus on generalizable characteristics that differentiate cancer data from normal tissue data. This is conceptually similar to forcing the model to view perturbations of the same image as similar, where the perturbation here is the natural variability among patients of the same class.

The pre-training step starts with an encoder (see Figure 1), which takes in unpaired pixel-level data from three classes: healthy, benign, and malignant. We apply a contrastive loss on the encoder embedding to train the network to cluster distinct classes. We optimize the total contrastive loss, $\mathcal{L}_{contr}$, which consists of *clustering* terms and *separation* terms:

$$\mathcal{L}_{contr} = \mathcal{L}_{clust} - \mathcal{L}_{sep} \tag{1}$$

where $\mathcal{L}_{clust}$ is the clustering loss we seek to minimize and $\mathcal{L}_{sep}$ is the separation loss we seek to maximize (hence, the negative sign in Equation (1)). We chose a summation of simple squared-error-based loss terms. While this strategy allows analysis of each loss term, we note that other contrastive losses may provide additional improvement. The clustering loss is given as follows:

$$\mathcal{L}_{clust} = \sum_{N_c} \alpha_c \sum_{N_{pospt}} (Emb_{anchor} - Emb_{positive})^2 \tag{2}$$

N_c refers to the number of classes (three in our case) and N_{pospt} refers to the number of patients with the same class as the anchor. The anchor within a single class, Emb_{anchor}, is the mean embedding of the class within the batch. An anchor-positive pair is formed by taking the mean embedding of each same-class patient in the batch ($Emb_{positive}$).

The separation loss is similar, but compares the anchor embedding to embeddings from the other classes:

$$\mathcal{L}_{sep} = \sum_{N_c} \beta_c \sum_{N_{negpt}} (Emb_{anchor} - Emb_{negative})^2 \tag{3}$$

where $Emb_{negative}$ is the mean embedding of the negative class patient within the batch. α_c and β_c denote the weights of the clustering and separation losses, respectively, for an anchor class c. We use an adaptive weighting of the losses, as described in Section 4.

3.3. Multitask Learning

Following contrastive pre-training, we add two task classifiers: a margin delineation classifier and a diagnosis classifier (see Figure 1). The margin delineation classifier labels pixels as lesion or healthy. For margin delineation, benign and malignant lesions are grouped in a general "lesion" class. The diagnosis classifier labels pixels as benign or malignant. During training, we group healthy samples with benign samples for the diagnosis task.

The total loss for the multitask learning phase ($\mathcal{L}_{MT}$) is as follows:

$$\mathcal{L}_{MT} = CE_{diag} + CE_{delin} + \mathcal{L}_{reg} \tag{4}$$

where CE_{diag} and CE_{delin} are the cross-entropy losses for the diagnosis and delineation classifiers, respectively. $\mathcal{L}_{reg}$ is a regularization loss that ensures consistency between the two classifiers. Specifically, the regularization loss is a cross-entropy loss that penalizes the network for labeling malignant samples differently. For example, we penalize cases when the margin delineation classifier labels a malignant sample as "healthy", but the diagnosis classifier labels the same malignant sample as "malignant"). Similarly, healthy tissue samples cannot simultaneously be cancerous, so any such predictions are penalized by the regularization loss.

3.4. Biological Basis and Imaging System

Though there are many types of cancers, Hanahan and Weinberg [31] initially described the development of cancer in 2000 as a collection of six basic cellular changes that underlie the development of most cancers. These six basic changes, or "hallmarks of cancer", were updated in 2011 to include changes in cellular metabolism as a potential additional hallmark [32]. In 2022, Hanahan [33] adopted "deregulating cellular metabolism" as an additional hallmark. As metabolic cofactors involved in oxidative phosphorylation and glycolysis [34–36], NADH and FAD have been investigated as potential autofluorescence biomarkers for pre-cancer or cancer by several studies (e.g., [37–39]).

Specifically for oral cancer, Sethupathi et al. [40] induced carcinogenesis using DMBA in a hamster model and measured the NADH and FAD autofluorescence. The autofluorescence values were used to calculate the redox ratio [40]. The redox ratio was compared through the development of cancer, showing that the redox ratio decreases as the induced lesions advance from normal to dysplasia to cancer [40]. In human cell lines, Shah et al. [41] show increased NADH and FAD autofluorescence in cancerous cell lines compared to normal cells with p-values < 0.05. Based on such insights, the imaging system excites tissue in a single excitation band and collects the response from endogenous fluorophores to quantify the optical characteristics of the tissue.

The collected data capture both the intensity of the response, as well as the temporal dynamics as the fluorescence decays over time. The tissue autofluorescence is collected in three emission bands (channels) corresponding to collagen, reduced *nicotinamide adenine dinucleotide* (NADH), and *flavin adenine dinucleotide* (FAD). The optical properties of collagen describe the structural characteristics of the tissue, which have been shown to change with neoplastic transformation (e.g., extracellular matrix remodeling) [42]. NADH and FAD autofluorescence can be used to calculate the optical redox ratio, which describes the metabolic state of the tissue [43–46]. A lower redox ratio is associated with an increase in cellular metabolism, which is an indicator of malignant transformation [47]. In head and neck cancers, changes in tissue concentration of NADH and FAD have been demonstrated [41,48,49]. The lifetime of the collected autofluorescence provides additional information about the tissue microenvironment, including protein binding changes associated with cancer development [37,50].

Further design details for the imaging system used for data collection have been previously reported by Cheng et al. [51]. The temporal resolution is 0.25 ns and the *field of view* (FOV) is a circular region with an approximate diameter of 1 cm. The system is a single-excitation system, with a 355 nm excitation laser that exposes the tissue to only 2.8 mJ of energy. The amount of energy deposited to the tissue is within the *maximum permissible exposure* (MPE) of 29.8 mJ set by the *American National Standards Institute* (ANSI) [52]. The pulse width of the excitation laser is 1ns. The resulting autofluorescence emission is captured in three emission bands. The first band is 390 $\pm$ 20 nm to capture collagen autofluorescence. The second band is 452 $\pm$ 22.5 nm to capture NADH autofluorescence. The final emission band is >500 nm to capture FAD autofluorescence. The time required to image is less than 3 s for each image. For each patient, we collect an image of the center of the oral lesion and an image of normal-appearing tissue. We non-invasively image each patient's oral lesion before the biopsy. Following our imaging protocol, patients undergo biopsies that are used to generate a pathology diagnosis for our ground-truth labels. The raw data collected by the system is a biexponential-like fluorescence decay for each channel at each pixel location (see Figure 1F).

3.5. Data and Preprocessing Steps

Our dataset consists of 67 lesion images and 67 healthy images, with an image size of 160 $\times$ 160 pixels (approx. 3.4 million pixels). The lesion diagnoses are as follows: 33 benign, 5 dysplasia, and 29 *squamous cell carcinoma* (SCC). For the classification tasks, we combined dysplasia and SCC in the same group. Institutional review board approval for the study was provided by Hamad Medical Corporation (Doha, Qatar). A detailed breakdown of the lesion distribution by diagnosis and anatomical location is given in Table 1. Notably, the dataset contains imbalance at both the diagnosis level (only five cases are diagnosed as dysplasia) and at the anatomical location level (over half of the lesions are located in the mucosa or tongue regions).

Cross-validation folds for a single trial were generated by randomly splitting the images into 10 folds, preserving class balance as much as possible within each fold. Within each run, we used 7 folds for the train set, 2 folds for the development set, and 1 fold for the test set. We conducted ten runs for each trial, ensuring that each fold was used as the testing set once per trial. We refer to the average results of all ten runs as a trial. We repeated the random splits 10 times, resulting in a total of 10 trials (10 trials $\times$ 10 runs per trial = 100 models).

Table 1. Lesion distribution in the database.

Location	Benign	Dysplasia	SCC
Mucosa	10	3	9
Floor of Mouth	2	0	1
Gingiva	0	2	3
Lip	10	0	2
Mandible	0	0	1
Palate	1	0	0
Maxilla	0	0	1
Retromolar	1	0	0
Tongue	9	0	12
Total	33	5	29

Each image was pre-processed using median filtering with a sliding window across pixels to increase the SNR. Median filtering with a sliding window increased the SNR but retained the original image size (160 $\times$ 160 pixels). After median filtering, pixels with low SNR were masked to exclude from training. At the pixel level, the first preprocessing step involved signal inversion for each decay. Next, each channel was zero-padded to 300 samples (total network input of 3 channels $\times$ 300 samples per channel = 900-dimensional

input). A calibration factor was applied to each channel to adjust for day-to-day variations in the system. Finally, each pixel signal (concatenated decays of the three channels) was normalized to sum to 100 to adjust for the different gains used during data collection. Deconvolution of the instrument response from the raw fluorescence decay, as typically performed in FLIM data analysis, was not needed in the proposed framework. An example of a pre-processed 3-channel decay, ready for input to the network, is shown in Figure 1 (example pixel).

3.6. Implementation

The specific structures for the encoder and classifiers are given in Figure 2. Each layer has three components: a fully connected layer, internal regularization, and activation. We used batch normalization in the encoder for regularization. Batch normalization has been reported to smooth the training surface [53], speed training [53,54], and improve generalization [53,54]. As described in Section 4, we also found that batch normalization stabilizes the separation terms in our contrastive loss. We trained the encoder with the contrastive loss given in Equation (1). We applied the loss to the fourth encoder layer after batch normalization and before the *rectified linear unit* (ReLU) activation. We optimized with Adam [55] (learning rate of 1×10^{-5}, batch size of 512). For additional stabilization, we used gradient clipping at 0.25. We trained the encoder for 10 epochs and picked the best model based on the development set. After the encoder was trained, we added the task classifiers.

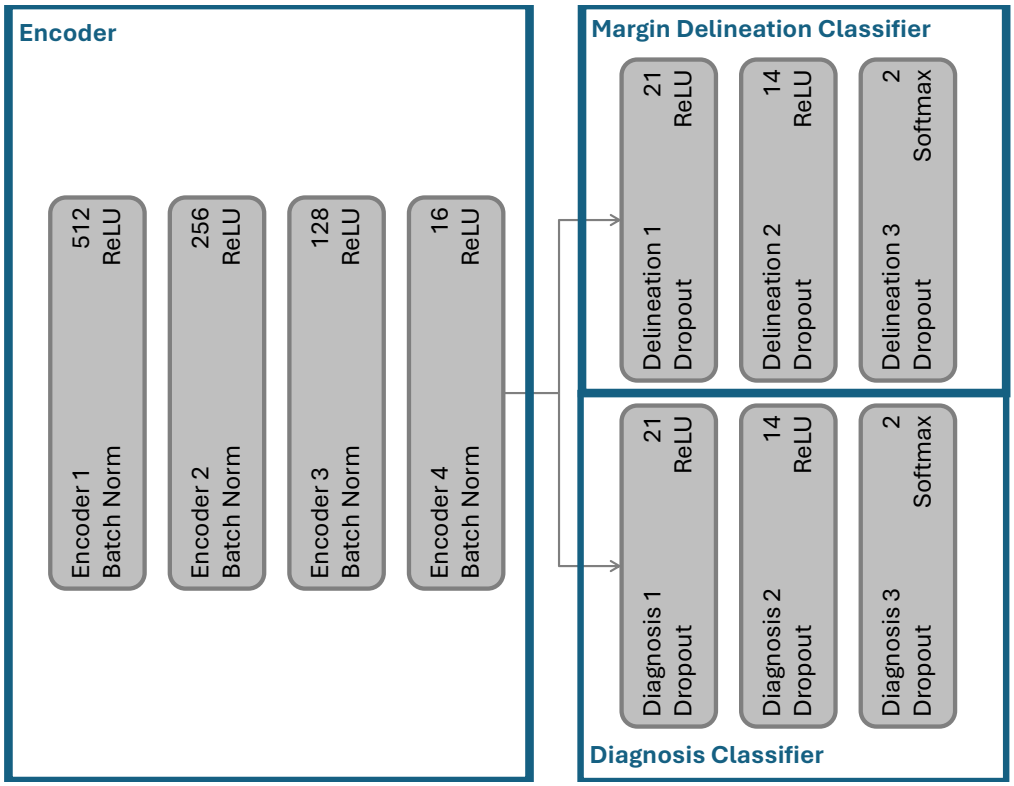


Figure 2. The network structure of the proposed architecture.

In the task classifiers, we used dropout for regularization ($p = 0.5$). By randomly removing nodes during training, dropout resulted in an ensemble-like model and reduced overfitting [56]. We optimized Equation (4) using Adam [55] (learning rate of 1×10^{-5}, batch size 256). The number of trainable parameters was 630,814 and the number of *floating point operations* (FLOPS) was 0.00126 G. The model was trained for an additional five epochs, with the best model selected based on the development set's performance. Due to class

imbalance (heavily biased toward the healthy class), we used sample weights calculated based on the training distribution, employing the *class weight* from *sci-kit learn* [57].

The task classifiers are trained on the pixel level. However, we report results at the image level. To determine the image-level label, we aggregate the labels of each pixel from an image and use a 50% threshold, taking the majority label of the image's pixels.

4. Results

The experiments in this section are organized as follows: contrastive loss function validation (Section 4.1), contrastive encoder training (Section 4.2), full model evaluation (Sections 4.3 and 4.4), and analysis of the contributions of the model's components (Section 4.5).

4.1. Loss Function Validation

We first used a toy example to understand how our loss function affected clustering (see Figure 3). As we developed our method, the synthetic data functioned as a simplistic simulation, where we had better control over the difficulty of the clustering task compared with our actual data. In the actual data, we could not control noise levels or the distances between classes. By gradually increasing the difficulty in synthetic data (starting with two classes and moving to three classes), we learned to adaptively weight the loss components and use batch normalization to achieve better clustering separation and keep the loss components from diverging. We used the *make moons* tool from *sci-kit learn* [57] to generate synthetic data and used dense layers with the contrastive loss function. Starting with a two-class example (see Figure 3a), we can show that the loss function clusters each class together, away from the other class (Figure 3b). Figure 3 also shows the results of each component of the loss function. As expected, the clustering component effectively clustered a single class, as seen in Figure 3c,d. The separation loss, however, presented more interesting behavior. Without batch normalization, the separation loss diverged (Figure 3e,f). If we added a batch normalization layer, the separation loss decreased in a controlled manner, and the classes separated as expected (Figure 3g,h).

Next, we introduced a third class of synthetic data to determine if our loss function could effectively separate three classes (Figure 4a). As shown in Figure 4b, only two classes separated well when the loss components had a static weight set at the beginning of the training. The training losses in Figure 5a show that one of the separation loss components began to dominate, resulting in good separation between the purple class and the other two. However, because the model received significant rewards for distancing the purple class from the other two, the separation of the other two classes was ignored. To balance the losses, we adaptively weighted the separation losses (β_c from Equation (3)) inversely proportional to the distance between the class means:

$$\beta_c = \frac{1}{dist_{A,Neg}} \tag{5}$$

where $dist_{A,Neg}$ is the squared distance between the means of the anchor class and the negative class. The weights are normalized such that they sum to one. The resulting weighting for a loss component is larger if two classes are poorly separated, but decreases as the separation increases. As shown in Figure 4c, the adaptive weighting strategy results in separated classes with nearly equidistant cluster centers. In addition, the loss curves in Figure 5b show that none of the separation loss components dominates.

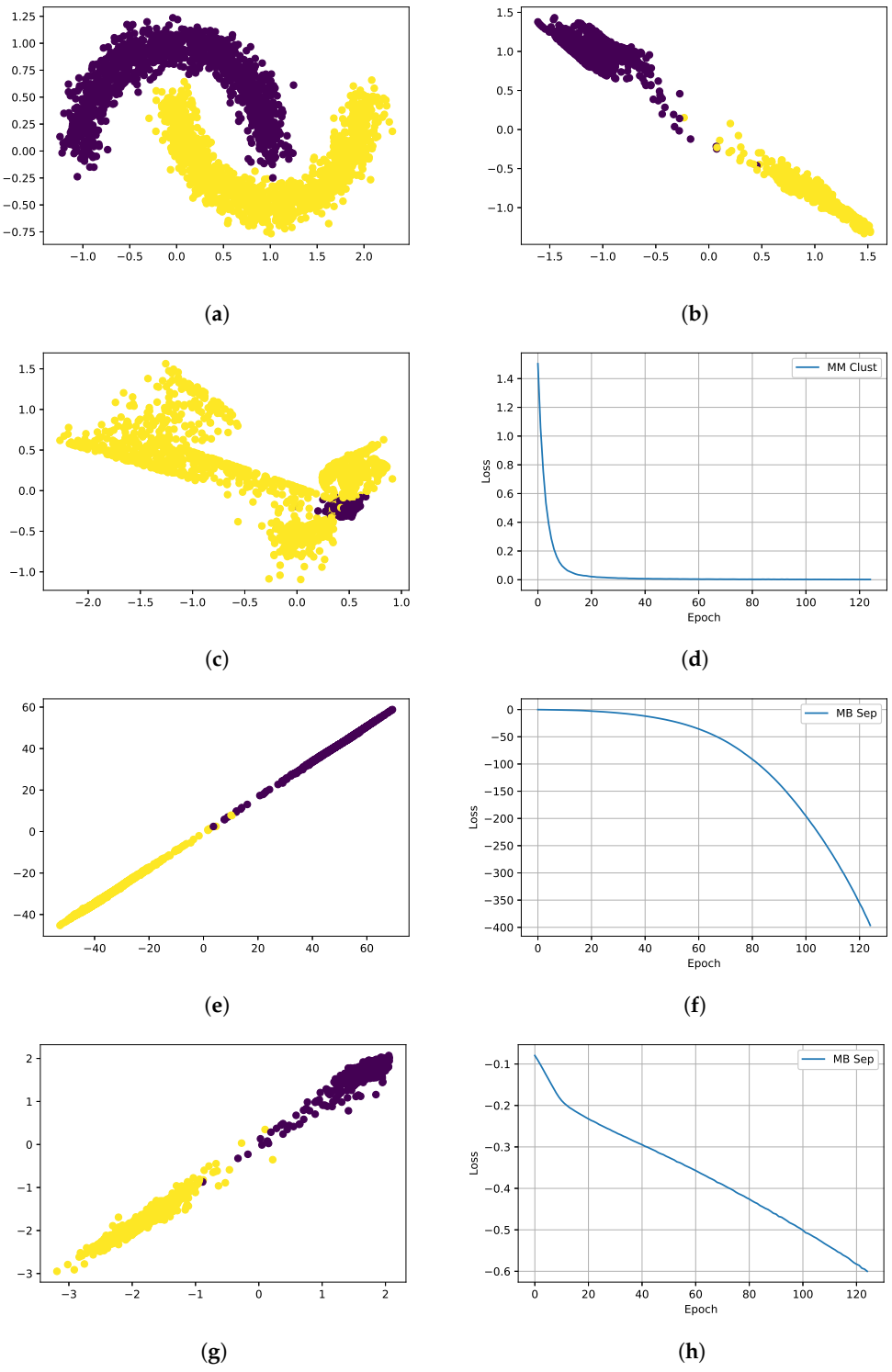
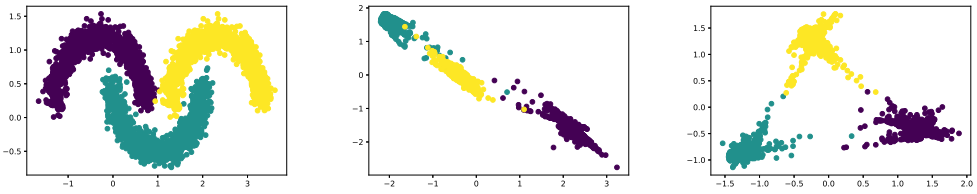
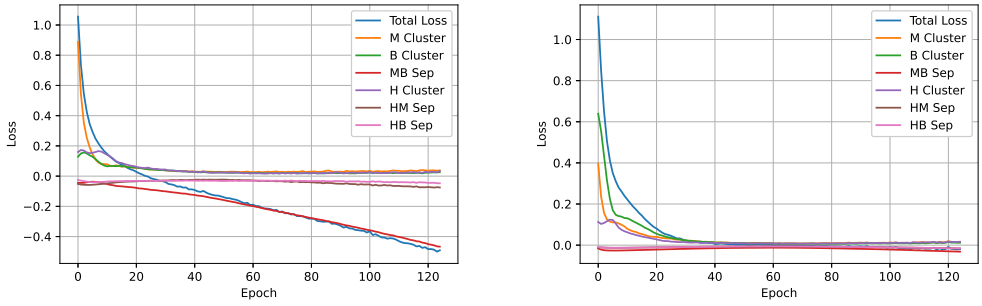


Figure 3. Contrastive training with synthetic data. The clustering and separation losses work together to separate the classes. Each component functions alone as expected when batch normalization is used. (**a**) Synthetic 2 class training data. (**b**) After: Clust. and Sep. losses with batch norm. (**c**) Single class clustering result. (**d**) Single class clustering loss. (**e**) Sep. loss collapse w/o batch norm. (**f**) Sep. loss training curve w/o batch norm. (**g**) Sep. loss only training w/batch norm. (**h**) Separation loss curve with batch norm.



(**a**) (**b**) (**c**)

Figure 4. Contrastive training with three classes. All three classes were separable after using the adaptive weighting scheme. (**a**) Synthetic 3 class training data. (**b**) Final result using clustering and separation losses without adaptive weighting. (**c**) Final result using clustering and separation losses with adaptive weighting.



(**a**) (**b**)

Figure 5. Loss functions for contrastive training with three classes. Each component of the loss function is more stable with the adaptive weighting. (**a**) Training losses without adaptive weighting. (**b**) Training losses with adaptive weighting.

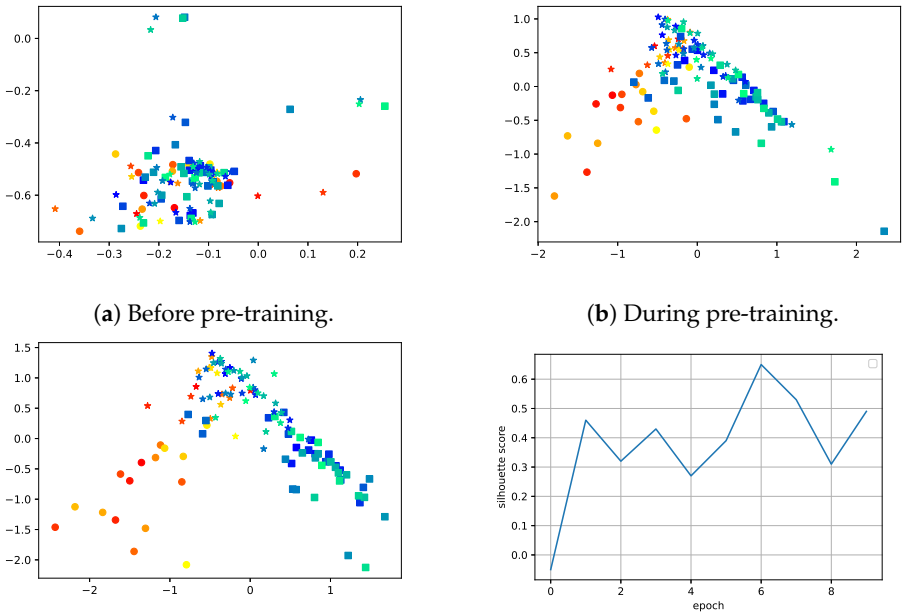
4.2. maFLIM Contrastive Pre-Training

To visualize the progress of contrastive pre-training on maFLIM data, we used the encoder and added a two-dimensional layer on which the contrastive loss operated (i.e., the dimension of the bottleneck representation was reduced from 16 to 2 for visualization). We also monitored the silhouette score [58] of the two-dimensional representation using the *sci-kit learn* implementation [57]. The silhouette score quantified the clustering performance and can be calculated as follows:

$$S = \frac{B - A}{max(A, B)} \tag{6}$$

where A is the mean distance between a sample and all other points in the same class and B is the mean distance between a sample and all other points in the next nearest cluster. The range of the silhouette score is from -1 to 1. A negative value indicates wrong assignments and a large positive value indicates high intra-class clustering and high inter-class separation. We use the Euclidean distance as the distance metric, although any distance metric may be used.

Figure 6 shows the 2D representations through contrastive training with the corresponding silhouette score. The warm colors (yellows and oranges) represent the mean embedding from patients with malignant lesions. The cool colors (blues and greens) represent the mean embedding from patients with benign lesions. The different shades in the colors represent different patients. The circles, squares, and stars represent the mean embedding from malignant, benign, and healthy images, respectively. For example, a patient with a malignant lesion will have two points plotted on the graph: an orange or yellow circle and a star in the same color.



(**a**) Before pre-training. (**b**) During pre-training.

(**c**) After pre-training. (**d**) Silhouette score during pre-training.

Figure 6. Two-dimensional (2D) contrastive feature space representation and silhouette scores during pre-training. The shapes represent each class. The stars represent healthy samples. The circles represent malignant samples. The squares represent benign samples. The cool colors represent samples from patients with benign lesions, while the warm colors represent samples from patients with malignant lesions. The shades of warm and cool colors help differentiate between tightly clustered patients, but have no additional meaning.

Using these two-dimensional plots, we found that the weights for Equation (2) could cause the representation space to collapse, resulting in all points clustering together, regardless of their class (i.e., the axes in the plots would span very little space). To prevent the space from collapsing, we implemented an adaptive weighting scheme for the α_c values based on the silhouette score. As the silhouette score increased, indicating good clustering, we gradually decreased the α_c weight. The adaptive weighting enabled the data to cluster quickly initially, but avoided instability due to dominance of the clustering term near the end of the contrastive pre-training.

Figure 6a shows the representation before contrastive training, where there was no separation between classes in two dimensions and the silhouette score was slightly negative. As training progressed (Figure 6b), the 2D representation began to show distinct class clustering and the silhouette score increased. As shown in Figure 6c, after training, the three classes were well clustered and the silhouette score neared 0.9. The contrastive pre-training step successfully groups each class and creates distance between different classes. In addition, Figure 6 shows that the silhouette score quantitatively describes the quality of the clusters visually depicted on the two-dimensional graphs. In the experiments that follow, we used the training set silhouette score to monitor training progress, as the 16-dimensional embedding in the full model cannot be directly visualized.

4.3. Diagnosis

We report the results of the full model (contrastive encoder plus multitask classification) on the diagnosis task in Table 2. Although the diagnosis classifier was trained using benign, malignant, and healthy samples, the evaluation was conducted only on the lesion samples. We compare the full model to four baselines: SVM with SFS, SVM with L1 regularization, an autoencoder model, and a multitask model. We have previously reported results with the SVM and autoencoder methods [13], while the multitask baseline enabled a more direct evaluation of the contrastive pre-training step. The multitask baseline

refers to the full model without the contrastive pre-training step. On average, adding the contrastive pre-training step improved the test performance by 2.33%, emphasizing the effectiveness of our patient normalization step. In comparison to the autoencoder, the full model showed an average improvement of 1.46%. In addition, the full model's increase of 8.34% in specificity represents a statistically significant improvement over the autoencoder method, as determined by a paired, one-tailed *t*-test ($p = 0.0070$). We note that the full model achieved better performance despite training on fewer lesion images than the autoencoder method, which used training images from two domains. When we compare this approach with the single-task SVM baseline proposed in Caughlin et al. [13], we observe improvements ranging from 2.75% to 4.87%. Using the average of sensitivity and specificity as a metric, a one-tailed, paired t-test shows a significant improvement over both SVM methods. The *p*-value for the SVM SFS method was 0.0261, while the *p*-value for the SVM L1 method was 0.0452. Although the full model outperformed all single-task baselines, we cannot attribute the increase in performance solely to the multitask framework as the multitask-only baseline does not perform better than the autoencoder baseline. Rather, the combination of patient normalization and multitask training achieves the best performance. We present an ablation study of the model components for the diagnosis task in Table 3.

Table 2. Diagnosis results. Full: contrastive pre-training with adaptive weights followed by multitask learning with consistency loss. Multitask: no contrastive pre-training. AE: autoencoder and a single-task classifier method. SVM SFS: support vector machine with sequential feature selection. SVM L1: support vector machine with L1 regularization.

Model	Sens.	Spec.	Avg.	Prec.	F1	Acc.
Full	82.08	75.92	79.00	76.92	78.07	78.81
Multitask	77.92	75.42	76.67	76.02	74.77	76.50
AE	87.50	67.58	77.54	76.25	79.80	77.62
SVM SFS	81.00	67.25	74.13	73.82	74.92	74.05
SVM L1	79.17	73.33	76.25	78.10	75.92	76.36

Table 3. Ablation results: diagnosis task. PT: pre-train. ACW: adaptive clustering weighting. CL: consistency loss. MT: multitask learning. Avg.: average of sensitivity and specificity.

Task	Clust. PT	Sep. PT	ACW	CL	MT	Sens.	Spec.	Avg.
	✓	✓	✓	✓	✓	82.08	75.92	79.00
	✓	✓		✓	✓	77.58	76.00	76.79
	✓	✓	✓			74.25	79.58	76.92
Diag.					✓	76.50	75.00	75.75
				✓	✓	77.92	75.42	76.67
	✓		✓	✓	✓	83.00	73.58	78.29
		✓		✓	✓	79.50	75.50	77.50
						71.92	81.50	76.71

4.4. Margin Delineation

We evaluate the margin delineation classifier for two clinical use cases. The first case is margin delineation between malignant and healthy tissue. This classifier would be used when a patient has a known malignant lesion. The classifier would aid resection by showing the clinician how much tissue to remove. The second case involves margin delineation between a lesion (regardless of diagnosis) and healthy tissue. This classifier would be used when a patient presents a suspicious lesion but has not undergone a biopsy or received a confirmed diagnosis. By removing the entire lesion and ensuring the boundary is also excised, the patient may not need to return for further resection, even if the diagnosis is malignant. We anticipate that this task will be more challenging due to the difficulty in distinguishing between benign and healthy tissue samples.

As in the diagnosis case, the contrastive pre-training plus multitask framework (full model) outperforms the multitask-only network by 0.6% to 3.4%. In addition, the full model has good performance on both types of margin delineation, reducing the performance gap between use cases from 6.17% to 3.37%. Although margin delineation is an auxiliary task in our framework, Table 4 shows that the classifier achieves good performance on margin delineation as well.

Table 4. Ablation results: margin delineation task. PT: pre-train. ACW: adaptive clustering weighting. CL: consistency loss. MT: multitask learning. Avg.: average of sensitivity and specificity.

Task	Clust. PT	Sep. PT	ACW	CL	MT	Sens.	Spec.	Avg.
M-H	✓	✓	✓	✓	✓	91.83	79.31	85.57
	✓	✓		✓	✓	88.33	79.19	83.76
	✓	✓	✓			89.33	75.64	82.49
					✓	88.33	82.31	85.32
				✓	✓	84.83	85.10	84.97
	✓		✓	✓	✓	91.33	79.76	85.55
		✓		✓	✓	90.50	80.22	85.36
						89.83	80.98	85.41
L-H	✓	✓	✓	✓	✓	85.10	79.31	82.20
	✓	✓		✓	✓	79.00	79.22	79.11
	✓	✓	✓			84.02	75.64	79.83
					✓	75.36	82.31	78.83
				✓	✓	72.53	85.07	78.80
	✓		✓	✓	✓	80.72	79.76	80.24
		✓		✓	✓	80.24	80.22	80.23
						80.76	80.98	80.87

4.5. Analysis of the Model's Components

The performance of the network with various components removed (ablation study) is shown in Tables 3 and 4. Comparing rows 1 and 4 for each task, the tables show that the contrastive pre-training step improves average performance across all tasks. The consistency loss improves average diagnosis performance, our primary task, by 0.92% (see columns 2 and 3 in rows 4 and 5 of Table 3). When compared to single-task training, multitask learning produced similar results on the diagnosis task. However, the performance is more evenly balanced in terms of sensitivity and specificity in the multitask setting. Finally, using adaptive clustering weights in the contrastive step improves performance by 2.21% (columns 1 and 4 in rows 1 and 2 in Table 3).

5. Conclusions

Our multitask learning framework, enhanced with contrastive pre-training, improves diagnosis performance across all baselines and also delivers good performance on margin delineation as an auxiliary task. The contrastive step helps group the classes compactly together, away from the other classes. This strategy aids patient normalization, without requiring comparison to a reference sample for each patient at test time, and provides a favorable initialization for multitask training. The multitask classifiers allow the margin delineation task and diagnosis task training to benefit each other, further regularizing the network and improving generalization.

A limitation of our work involves the size of the dataset. However, we included several strategies to reduce potential problems with small data. Our pixel-level training process resulted in many more samples for training. Due to the novel data format, our models were trained at the pixel level rather than the image level. Since each image contained 160×160 pixels and we took images of each patient's normal and lesion tissues, we had $160 \times 160 \times 2 = 51{,}200$ pixels (samples) per patient before preprocessing. While we did exclude some of these pixels during preprocessing based on SNR, using 51,200 pixels per each of the 67 patients led to over 3.4 million pixels/samples. We note that our total patient

count is comparable to related research using fluorescence lifetime imaging systems (e.g., the work by Marsden et al. [6], with 53 patients compared to our 67 patients). In the related field of fluorescence lifetime imaging microscopy, several studies have used similar sample sizes. In a lung cancer application of fluorescence lifetime imaging microscopy, Wang et al. included a total of 31 patients in their study [59]. Ji et al. used a database of 71 patients with fluorescence lifetime imaging microscopy data for a machine learning application in cervical cancer risk [60].

In addition, we carefully designed our machine learning models to reduce bias in the predictions given the smaller sample sizes in our database. In our baselines and proposed methods, we used regularization methods to reduce overfitting and improve generalization. In the SVM baseline, we used L1 regularization. L1 regularization encourages some coefficients to go to zero, creating a simpler decision function that focuses on the most relevant features. In our proposed method, we implemented multitask learning as a form of regularization. In multitask learning, the network was constrained to find a solution that satisfied both tasks, leading to more robust representations and classifiers. We also limited the network size to reduce the likelihood of overfitting. Our feature extractor consists of only four layers, and the classifier has just three layers, including the final two-node classification layer. Finally, we meticulously applied a 10-fold cross-validation scheme and repeated the experiments for 10 full trials, ensuring that our results are an average over 100 different models.

Although our analysis only considers regularization and patient normalization through our contrastive pre-training, other automated medical tasks may benefit from similar approaches. For example, as datasets grow larger, our approach naturally extends to more specific classes. We may be able to extend our multitask setting to consider different grades of pre-cancer or to accommodate separate classes for different types of benign lesions. For example, given the adequate dataset size, our approach could be easily modified to provide a more specific diagnosis by adding additional clusters during pre-training that separate SCC, mild, moderate, and severe dysplasia. Following pre-training, the multitask learning framework could be modified to include additional outputs in the diagnosis classifier to accommodate the more specific classes of dysplasia and cancer. Using the contrastive approach may also improve domain generalization. Domain generalization is required for clinical translation when a trained model is deployed in a new location. Taking inspiration from the contrastive learning work in the natural image domain, (e.g., Kim et al. [61]), we would like to extend our work to the multi-center setting, where multiple small datasets from separate imaging centers need to be merged. In this setting, the clustering losses in the contrastive pre-training step would encourage samples from the same class to group compactly, regardless of domain shifts due to differences in imaging centers.

Author Contributions: Conceptualization, C.B. and J.A.J.; methodology, K.C., E.D.-S., S.C., R.C., B.A., J.J., M.M., M.A.-K., H.A.-E., J.A.J. and C.B.; software, K.C., E.D.-S. and R.C.; validation, J.A.J. and C.B.; formal analysis, K.C., E.D.-S., R.C., J.A.J. and C.B.; investigation, K.C., E.D.-S., R.C., J.A.J. and C.B.; resources, S.C., R.C., M.M., M.A.-K., H.A.-E., J.A.J. and C.B.; writing—original draft preparation, K.C. and C.B.; writing—review and editing, C.B. and J.A.J.; visualization, K.C.; supervision, C.B. and J.A.J.; project administration, J.A.J. and C.B.; funding acquisition, J.A.J. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Qatar National Resesrch Fund (NPRP8-1606-3-322), the Cancer Prevention and Research Institute of Texas (CPRIT grant RP180588), and the National Cancer Institute of the National Institutes of Health (NIH-NCI grant R01CA218739). Research reported in this publication was also supported in part by an Institutional Development Award (IDeA) from the NIH National Institute of General Medical Sciences (NIGMS) under grant number P20GM135009, and by the Oklahoma Tobacco Settlement Endowment Trust awarded to the University of Oklahoma, Stephenson Cancer Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Oklahoma Tobacco Settlement Endowment Trust.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Hamad Medical Corporation (Doha, Qatar) HMC-IRB 16332/16 with approval date of 29 January 2019.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Please contact Javier Jo (javierjo@ou.edu) regarding data access. As the data used in this study were obtained from human subjects, data are not publicly available.

Conflicts of Interest: Authors Mathias Martinez, Moustafa Al-Khalil and Hussain Al-Enazi were employed by the company Hamad Medical Corporation. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
3. Welikala, R.A.; Remagnino, P.; Lim, J.H.; Chan, C.S.; Rajendran, S.; Kallarakkal, T.G.; Zain, R.B.; Jayasinghe, R.D.; Rimal, J.; Kerr, A.R.; et al. Fine-tuning deep learning architectures for early detection of oral cancer. In Proceedings of the International Symposium on Mathematical and Computational Oncology, Virtual, 8–10 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 25–31.
4. Zhang, H.; Li, W.; Zhang, H. An Image Recognition Framework for Oral Cancer Cells. *J. Healthc. Eng.* **2021**, *2021*, 2449128. [CrossRef]
5. Lu, J.; Sladoje, N.; Runow Stark, C.; Darai Ramqvist, E.; Hirsch, J.M.; Lindblad, J. A deep learning based pipeline for efficient oral cancer screening on whole slide images. In Proceedings of the International Conference on Image Analysis and Recognition, Póvoa de Varzim, Portugal, 24–26 June 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 249–261.
6. Marsden, M.; Weyers, B.W.; Bec, J.; Sun, T.; Gandour-Edwards, R.F.; Birkeland, A.C.; Abouyared, M.; Bewley, A.F.; Farwell, D.G.; Marcu, L. Intraoperative Margin Assessment in Oral and Oropharyngeal Cancer Using Label-Free Fluorescence Lifetime Imaging and Machine Learning. *IEEE Trans. Biomed. Eng.* **2021**, *68*, 857–868. [CrossRef] [PubMed]
7. Jiang, X.; Wu, J.; Wang, J.; Huang, R. Tobacco and oral squamous cell carcinoma: A review of carcinogenic pathways. *Tob. Induc. Dis.* **2019**, *17*, 29. [CrossRef]
8. Jo, J.; Cheng, S.; Cuenca-Martinez, R.; Duran-Sierra, E.; Malik, B.; Ahmed, B.; Maitland, K.; Cheng, Y.S.; Wright, J.; Reese, T. Endogenous Fluorescence Lifetime Imaging (FLIM) Endoscopy For Early Detection Of Oral Cancer And Dysplasia. In Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2018), Honolulu, HI, USA, 18–21 July 2018; pp. 3009–3012. [CrossRef]
9. Vasanthakumari, P.; Romano, R.; Rosa, R.; Salvio, A.; Yakovlev, V.; Kurachi, C.; Jo, J. Classification of skin-cancer lesions based on fluorescence lifetime imaging. In Proceedings of the SPIE Medical Imaging, 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, Houston, TX, USA, 18–20 February 2020; Volume 11317. [CrossRef]
10. Cheng, J.Y.; Goh, H.; Dogrusoz, K.; Tuzel, O.; Azemi, E. Subject-aware contrastive learning for biosignals. *arXiv* **2020**, arXiv:2007.04871.
11. Duran-Sierra, E.; Cheng, S.; Cuenca-Martinez, R.; Malik, B.; Maitland, K.C.; Cheng, Y.L.; Wright, J.; Ahmed, B.; Ji, J.; Martinez, M.; et al. Clinical label-free biochemical and metabolic fluorescence lifetime endoscopic imaging of precancerous and cancerous oral lesions. *Oral Oncol.* **2020**, *105*, 104635. [CrossRef]
12. Duran-Sierra, E.; Cheng, S.; Cuenca, R.; Ahmed, B.; Ji, J.; Yakovlev, V.V.; Martinez, M.; Al-Khalil, M.; Al-Enazi, H.; Cheng, Y.S.L.; et al. Machine-Learning Assisted Discrimination of Precancerous and Cancerous from Healthy Oral Tissue Based on Multispectral Autofluorescence Lifetime Imaging Endoscopy. *Cancers* **2021**, *13*, 4751. [CrossRef]
13. Caughlin, K.; Duran-Sierra, E.; Cheng, S.; Cuenca, R.; Ahmed, B.; Ji, J.; Yakovlev, V.; Martinez, M.; Al-Khalil, M.; Al-Enazi, H.; et al. End-to-End Neural Network for Feature Extraction and Cancer Diagnosis of In Vivo Fluorescence Lifetime Images of Oral Lesions. In Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2021), Guadalajara, Mexico, 1–5 November 2021.
14. Weinberger, K.Q.; Blitzer, J.; Saul, L. Distance metric learning for large margin nearest neighbor classification. *Adv. Neural Inf. Process. Syst.* **2005**, *18*, 1473–1480.
15. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, MA, USA, 7–12 June 2015; pp. 815–823. [CrossRef]
16. Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1–9.

17. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
18. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 539–546.
19. Goswami, S.; Mehta, S.; Sahrawat, D.; Gupta, A.; Gupta, R. Heterogeneity loss to handle intersubject and intrasubject variability in cancer. *arXiv* **2020**, arXiv:2003.03295.
20. Choudhary, A.; Wu, H.; Tong, L.; Wang, M.D. Learning to evaluate color similarity for histopathology images using triplet networks. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Niagara Falls, NY, USA, 7–10 September 2019; pp. 466–474.
21. Vu, Y.N.T.; Wang, R.; Balachandar, N.; Liu, C.; Ng, A.Y.; Rajpurkar, P. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In Proceedings of the Machine Learning for Healthcare Conference, PMLR, Virtual, 6–7 August 2021; pp. 755–769.
22. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
23. Seo, H.; Yu, L.; Ren, H.; Li, X.; Shen, L.; Xing, L. Deep neural network with consistency regularization of multi-output channels for improved tumor detection and delineation. *IEEE Trans. Med Imaging* **2021**, *40*, 3369–3378. [CrossRef]
24. Khosravan, N.; Bagci, U. Semi-supervised multi-task learning for lung cancer diagnosis. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 710–713.
25. Sainz de Cea, M.V.; Diedrich, K.; Bakalo, R.; Ness, L.; Richmond, D. Multi-task learning for detection and classification of cancer in screening mammography. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 241–250.
26. Hu, H.; Wang, X.; Zhang, Y.; Chen, Q.; Guan, Q. A comprehensive survey on contrastive learning. *Neurocomputing* **2024**, *610*, 128645. [CrossRef]
27. Xu, S.; Zhang, X.; Wu, Y.; Wei, F. Sequence level contrastive learning for text summarization. In Proceedings of the AAAI Conference on Artificial Intelligence, Pomona, CA, USA, 24–28 October 2022; Volume 36, p. 11556.
28. You, Y.; Chen, T.; Shen, Y.; Wang, Z. Graph contrastive learning automated. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 12121–12132.
29. Kuang, H.; Zhu, Y.; Zhang, Z.; Li, X.; Tighe, J.; Schwertfeger, S.; Stachniss, C.; Li, M. Video contrastive learning with global context. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3195–3204.
30. Fernando, T.; Gammulle, H.; Denman, S.; Sridharan, S.; Fookes, C. Deep learning for medical anomaly detection–a survey. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–37. [CrossRef]
31. Hanahan, D.; Weinberg, R.A. The hallmarks of cancer. *Cell* **2000**, *100*, 57–70. [CrossRef]
32. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef]
33. Hanahan, D. Hallmarks of cancer: New dimensions. *Cancer Discov.* **2022**, *12*, 31–46. [CrossRef]
34. Croce, A.C.; Bottiroli, G. Autofluorescence spectroscopy and imaging: A tool for biomedical research and diagnosis. *Eur. J. Histochem. EJH* **2014**, *58*, 2461. [CrossRef]
35. Bartolomé, F.; Abramov, A.Y. Measurement of mitochondrial NADH and FAD autofluorescence in live cells. *Mitochondrial Med. Vol. I Probing Mitochondrial Funct.* **2015**, *1264*, 263–270.
36. Kolenc, O.I.; Quinn, K.P. Evaluating cell metabolism through autofluorescence imaging of NAD (P) H and FAD. *Antioxidants Redox Signal.* **2019**, *30*, 875–889. [CrossRef] [PubMed]
37. Skala, M.C.; Riching, K.M.; Gendron-Fitzpatrick, A.; Eickhoff, J.; Eliceiri, K.W.; White, J.G.; Ramanujam, N. In vivo multiphoton microscopy of NADH and FAD redox states, fluorescence lifetimes, and cellular morphology in precancerous epithelia. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19494–19499. [CrossRef]
38. Ramanujan, V.K.; Zhang, J.H.; Biener, E.; Herman, B. Multiphoton fluorescence lifetime contrast in deep tissue imaging: Prospects in redox imaging and disease diagnosis. *J. Biomed. Opt.* **2005**, *10*, 051407. [CrossRef]
39. Meier, J.D.; Xie, H.; Sun, Y.; Sun, Y.; Hatami, N.; Poirier, B.; Marcu, L.; Farwell, D.G. Time-resolved laser-induced fluorescence spectroscopy as a diagnostic instrument in head and neck carcinoma. *Otolaryngol. Neck Surg.* **2010**, *142*, 838–844. [CrossRef] [PubMed]
40. Sethupathi, R.; Gurushankar, K.; Krishnakumar, N. Optical redox ratio differentiates early tissue transformations in DMBA-induced hamster oral carcinogenesis based on autofluorescence spectroscopy coupled with multivariate analysis. *Laser Phys.* **2016**, *26*, 116202. [CrossRef]
41. Shah, A.T.; Demory Beckler, M.; Walsh, A.J.; Jones, W.P.; Pohlmann, P.R.; Skala, M.C. Optical metabolic imaging of treatment response in human head and neck squamous cell carcinoma. *PLoS ONE* **2014**, *9*, e90746. [CrossRef] [PubMed]
42. Pavlova, I.; Williams, M.; El-Naggar, A.; Richards-Kortum, R.; Gillenwater, A. Understanding the biological basis of autofluorescence imaging for oral cancer detection: High-resolution fluorescence microscopy in viable tissue. *Clin. Cancer Res.* **2008**, *14*, 2396–2404. [CrossRef]
43. Gulledge, C.; Dewhirst, M. Tumor oxygenation: A matter of supply and demand. *Anticancer Res.* **1996**, *16*, 741–749.

44. Drezek, R.; Brookner, C.; Pavlova, I.; Boiko, I.; Malpica, A.; Lotan, R.; Follen, M.; Richards-Kortum, R. Autofluorescence microscopy of fresh cervical-tissue sections reveals alterations in tissue biochemistry with dysplasia. *Photochem. Photobiol.* **2001**, *73*, 636–641. [CrossRef]
45. Ramanujam, N.; Richards-Kortum, R.; Thomsen, S.; Mahadevan-Jansen, A.; Follen, M.; Chance, B. Low temperature fluorescence imaging of freeze-trapped human cervical tissues. *Opt. Express* **2001**, *8*, 335–343. [CrossRef]
46. Chance, B.; Schoener, B.; Oshino, R.; Itshak, F.; Nakase, Y. Oxidation-reduction ratio studies of mitochondria in freeze-trapped samples. NADH and flavoprotein fluorescence signals. *J. Biol. Chem.* **1979**, *254*, 4764–4771. [CrossRef] [PubMed]
47. Zhang, Z.; Blessington, D.; Li, H.; Busch, T.M.; Glickson, J.D.; Luo, Q.; Chance, B.; Zheng, G. Redox ratio of mitochondria as an indicator for the response of photodynamic therapy. *J. Biomed. Opt.* **2004**, *9*, 772–778. [CrossRef] [PubMed]
48. Müller, M.G.; Valdez, T.A.; Georgakoudi, I.; Backman, V.; Fuentes, C.; Kabani, S.; Laver, N.; Wang, Z.; Boone, C.W.; Dasari, R.R.; et al. Spectroscopic detection and evaluation of morphologic and biochemical changes in early human oral carcinoma. *Cancer Interdiscip. Int. J. Am. Cancer Soc.* **2003**, *97*, 1681–1692. [CrossRef]
49. Jo, J.A.; Applegate, B.E.; Park, J.; Shrestha, S.; Pande, P.; Gimenez-Conti, I.B.; Brandon, J.L. In vivo simultaneous morphological and biochemical optical imaging of oral epithelial cancer. *IEEE Trans. Biomed. Eng.* **2010**, *57*, 2596–2599. [CrossRef] [PubMed]
50. Shah, A.T.; Heaster, T.M.; Skala, M.C. Metabolic imaging of head and neck cancer organoids. *PLoS ONE* **2017**, *12*, e0170415. [CrossRef] [PubMed]
51. Cheng, S.; Cuenca, R.; Liu, B.; Malik, B.; Jabbour, J.; Maitland, K.; Wright, J.; Cheng, Y.S.; Jo, J. Handheld multispectral fluorescence lifetime imaging system for in vivo applications. *Biomed. Opt. Express* **2014**, *5*, 921–931. [CrossRef]
52. *ANSI Z136.1–2007* ; Safe Use of Lasers. American National Standards Institute: Orlando, FL, USA, 2007.
53. Santurkar, S.; Tsipras, D.; Ilyas, A.; Madry, A. How does batch normalization help optimization? In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, BC, Canada, 3–8 December 2018; pp. 2488–2498.
54. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning (PMLR 2015), Lille, France, 7–9 July 2015; Volume 37, pp. 448–456.
55. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–13.
56. Srivastava, N.; Hinton, G.E.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
57. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
58. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
59. Wang, M.; Tang, F.; Pan, X.; Yao, L.; Wang, X.; Jing, Y.; Ma, J.; Wang, G.; Mi, L. Rapid diagnosis and intraoperative margin assessment of human lung cancer with fluorescence lifetime imaging microscopy. *BBA Clin.* **2017**, *8*, 7–13. [CrossRef]
60. Ji, M.; Zhong, J.; Xue, R.; Su, W.; Kong, Y.; Fei, Y.; Ma, J.; Wang, Y.; Mi, L. Early detection of cervical cancer by fluorescence lifetime imaging microscopy combined with unsupervised machine learning. *Int. J. Mol. Sci.* **2022**, *23*, 11476. [CrossRef] [PubMed]
61. Kim, D.; Yoo, Y.; Park, S.; Kim, J.; Lee, J. Selfreg: Self-supervised contrastive regularization for domain generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9619–9628.