Back2Seg: Joint Background Estimation and Bacteria Segmentation on Optical Endomicroscopy Images

Back2Seg: Joint Background Estimation and Bacteria Segmentation on Optical Endomicroscopy Images

Mehmet Demirel[1], Bethany Mills[2], Erin Gaughan[2], Kevin Dhaliwal[2], and James R. Hopgood[1]

[1]IDCOM, School of Engineering, University of Edinburgh, Edinburgh, UK
[2]Edinburgh Medical School, University of Edinburgh, Edinburgh, UK

***Abstract*—Pneumonia, a lung infection typically caused by bacteria, requires swift and accurate diagnosis, especially in critical care. Optical endomicroscopy (OEM) facilitates real-time acquisition of in vivo and in situ optical biopsies, aiding in the quick identification of bacteria. However, the challenge of visually analyzing the vast number of images generated by the OEM in real-time can lead to delays in necessary treatments. To address this, we introduce Back2Seg, a novel approach for the segmentation of bacteria in OEM image sequences. Prior research mainly focused on exploiting bacteria motion or relied on less accurate unsupervised background estimation methods. In this regard, to enhance the background estimation and thus bacteria segmentation, Back2Seg employs a two-stage architecture with one sub-network dedicated to estimating the background using a Convolutional Neural Network (CNN)-Transformer architecture and the other is a dual-input network, processing both the original and the estimated background sequences to accurately segment the bacteria. Our experiments demonstrate that Back2Seg effectively integrates the advantages of both supervised and unsupervised learning techniques, showing a 4.62% increase in correlation with annotations over unsupervised models and a 1.05 reduction in root mean squared error (RMSE), outperforming the top supervised approach.**

***Index Terms*—Bacteria Detection, Optical Endomicroscopy, CNN, Transformer, Background Estimation**

I. Introduction

Pneumonia, a lung infection often caused by bacteria, poses a significant risk, especially in intensive care settings, where its a leading cause of death [1], [2]. Effective treatment hinges on quickly starting antimicrobial therapy, which depends on fast identification and quantification of the bacteria in the lungs [3]. However, traditional imaging methods like X-rays and CT scans, mainly used to detect pneumonia, are slow, typically taking 48-72 hours to yield results [4], [5]. Optical endomicroscopy (OEM) offers a quicker alternative by providing real-time acquisition of in vivo and in situ optical biopsies, thus accelerating the diagnosis of specific medical conditions, including bacterial infections [5]–[7].

Nonetheless, the challenge lies in efficiently analyzing the large volume of images produced by OEM in real-time, which is particularly daunting for medical professionals who are faced with the manual visual inspection of these images. Such a process can introduce significant delays in diagnosing and initiating treatment. In this context, machine learning techniques can be utilized, offering a path to expedite the detection of bacteria within these images more swiftly and accurately, thereby streamlining the analysis process and reducing potential treatment delays.

Recently both unsupervised and supervised machine-learning techniques have been applied to detecting bacteria in OEM images. Most of the prior research mainly focused on unsupervised methods due to the annotated data limitations. These approaches include [8] and [9], which employed unsupervised methods for bacteria detection, where the input images are considered as a combination of intensity values corresponding to the background, contaminated by additive Gaussian noise, and possibly some sparse anomalies representing potential bacteria. [8] formulated the bacteria detection as a minimization problem, and estimated the unknown parameters using an alternating direction method of multipliers (ADMM). Conversely, [9] used a hierarchical Bayesian model (HBM) to detect bacteria in the images, utilizing Markov random fields (MRF) to describe the background. Nevertheless, both of these approaches have their own set of constraints. [8] required manual adjustment of parameters, whereas [9] represented bacteria as singular pixels (outliers), a model that does not accurately reflect the true nature of bacteria. To overcome these limitations, [10] proposed a fully automatic iterative HBM for detecting bacteria in OEM images. In this model, the image is considered to be a mixture of background, Gaussian noise, and additional anomalies representing bacteria. Within this framework, bacteria are depicted as objects with a circularly symmetric Gaussian shape, possessing an undefined full-width-at-half-maximum (FWHM). Although [10] presents a significant benchmark with considerable performance advantages, its high computational demands make it unsuitable for real-world applications. To overcome the speed limitations of the previous approaches, [11] proposed to generate synthetic OEM image sequences with bacteria, to train supervised machine learning approaches for real time analysis. In order to generate these synthetic image sequences, OEM lung image sequences without any bacteria are utilized as a background. To these healthy background image sequences, they introduce bacteria with two distinct motion models that were informed by medical professionals and by the physical nature of the OEM imaging system. Moreover, to effectively train with these synthetic images [11] proposed EmiNet, a two-stream Con-

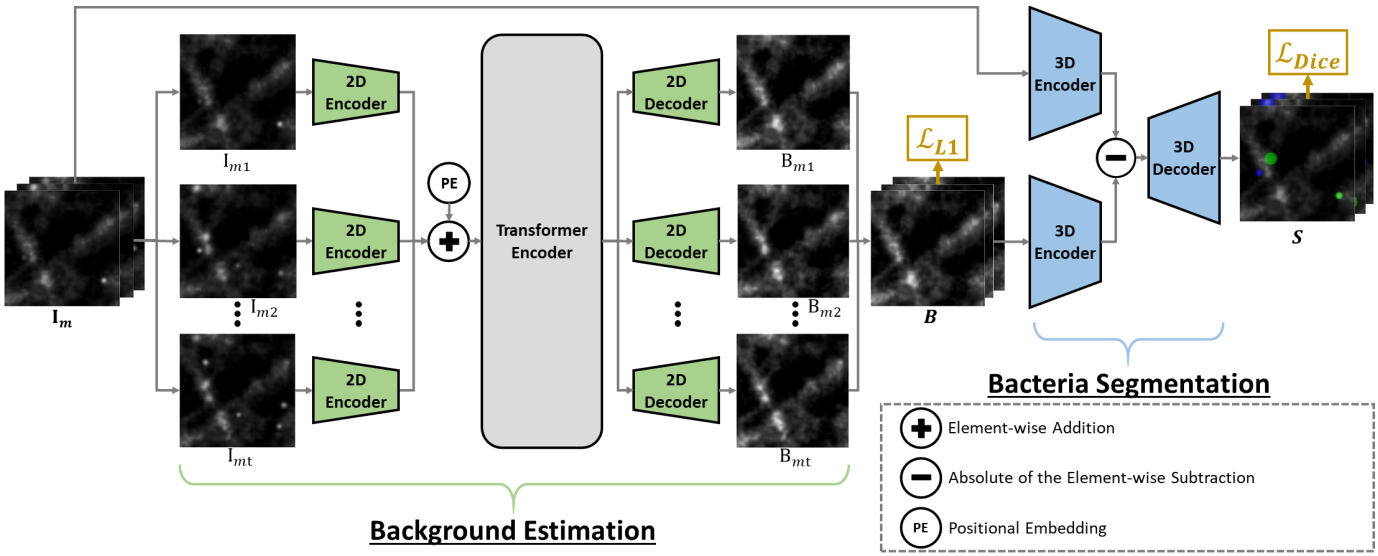

Fig. 1. An overview of the proposed Back2Seg model. The network consists of two stages. The first stage sub-network takes the OEM image sequence as input and estimates the background. The second stage takes the OEM image sequence and the estimated background sequence and segments the bacteria. Please note that skip connections are omitted in the figure to avoid clutter.

volutional Neural Network (CNN)-transformer-based encoder-decoder model. In this model one stream is used to capture the appearance features and the other stream captures the motion features. Training this model with the synthetic data improved the performance and allowed for real-time bacteria detection. In [11], the model focuses solely on the motion and appearance of bacteria for segmentation. Given that bacteria occupy a relatively small portion of the overall image, there is a risk that the model may incorrectly identify them as noise or background artifacts. To mitigate this, it would be advantageous to first perform a background estimation, which counts for the majority of the image. This background estimation could then be leveraged to enhance the segmentation of bacteria, similar to the unsupervised approaches [9], [10].

Therefore using the synthetic data proposed by [11], in this paper, we propose, Back2Seg, a two-stage model for segmenting bacteria in OEM image sequences. The network consists of a background estimation stage and a bacteria segmentation stage. The background estimation stage utilizes a hybrid of the CNN-Transformer model, digesting an OEM image sequence with bacteria, and aggregating temporal information to output the background image sequence without any bacteria. On the other hand, the bacteria segmentation stage takes the original image sequence and the estimated background image sequence as inputs and segments the bacteria in the image sequences.

To evaluate the performance of the proposed approach, experiments have been conducted on both real and synthetic datasets. Experiments show that the proposed background estimation approach outperforms existing methods described in Section IV-A. Moreover, we demonstrate that our method enhances the accuracy of detecting bacteria by reducing the root mean squared error (RMSE) between the number of detected bacteria and the manually counted bacteria, surpassing the performance of the current leading supervised detection technique and outperforming the correlation with the annotation counts of the state of the art unsupervised methods.

II. Method

An overview of the proposed Back2Seg model is shown in Fig. 1. The network consists of two stages. The first stage sub-network is designed for estimating the background of the OEM image sequences. It is composed of a hybrid architecture, with CNNs and Transformers. Specifically, CNNs are used to extract features from each image in the input sequence, with the Transformer built on top to aggregate spatial-temporal information and output framewise background estimation. The second stage sub-network is a two-stream 3D encoder-decoder network that is used to segment the bacteria in the image sequences by inputting the original image sequence and its estimated background image sequence.

A. Background Estimation

The OEM image sequences show slight movement due to breathing and as a result of this movement, the background is not stationary and shows changes and movement as well. Therefore, to capture these temporal variations in the background, the input images need to be processed together instead of individually estimating the background. In this regard, inspired by [12], we use a lightweight 2D CNN, $\mathbf{\Phi}_{Enc}$, that takes as input a sequence of images, $\mathbf{I}_m = \{I_{m_1}, I_{m_2}, \ldots, I_{m_t}\} \in \mathbb{R}^{t \times C_0 \times H_0 \times W_0}$ and the generates feature maps:

$$\{f_{m_1}, f_{m_2}, \ldots, f_{m_t}\} = \mathbf{\Phi}_{Enc}\left(\mathbf{I}_m\right)$$

where $f_i \in \mathbb{R}^{C_\Phi \times H_\Phi \times W_\Phi}$ refers to the feature map. H, W, C denote height, width, and channels.

These feature maps are then concatenated along the sequence and reshaped into a sequence of tokens. Following

[13], a learnable positional embedding is also added to keep track of the position of each token. Given by:

$$\mathbf{z}_0 = Reshape\left(\{f_{m_1}, f_{m_2}, \ldots, f_{m_t}\}\right) + \mathbf{E}_{pos}$$

where $\mathbf{z}_0 \in \mathbb{R}^{C_\Phi \times tH_\Phi W_\Phi}$ is the output feature embedding and $\mathbf{E}_{pos} \in \mathbb{R}^{C_\Phi \times tH_\Phi W_\Phi}$ is the position embedding.

Once we have the output feature embedding, it's inputted into the Transformer encoder [13] to capture the long-range spatial-temporal dependencies. The Transformer encoder is composed of L Transformer layers. Each layer consists of a Multi-Head Attention (MHA) block followed by Multi-Layer Perceptron (MLP) blocks [13]. Therefore the output of the l-th ($l \in [1, \ldots, L]$) transformer layer can be written as:

$$\acute{\mathbf{z}}_l = \text{MHA}(\text{LN}\left(\mathbf{z}_{l-1}\right)) + \mathbf{z}_{l-1},$$

$$\mathbf{z}_l = \text{MLP}(\text{LN}\left(\acute{\mathbf{z}}_l\right)) + \acute{\mathbf{z}}_l$$

where $LN(*)$ is the layer normalization, $\acute{\mathbf{z}}_l$ is the output of l-th Transformer layer and $\mathbf{z}_l$ is the Transformer encoder output.

For decoding, we reshape $\mathbf{z}_l$ into individual feature maps corresponding to each frame in the sequence. These feature maps are then passed through a lightweight 2D CNN decoder, $\mathbf{\Theta}_{Dec}$, with skip connections to recover high-resolution background estimations:

$$\mathbf{B} = \{B_{m_1}, B_{m_2}, \ldots, B_{m_t}\} = \mathbf{\Theta}_{Dec}\left(Reshape\left(\mathbf{z}_l\right)\right)$$

where $\mathbf{B} \in \mathbb{R}^{t \times C_0 \times H_0 \times W_0}$ is the estimated background.

The background estimation sub-network is trained using synthetic image sequences proposed in [11] using $L1$ loss $\mathcal{L}_{L1}$:

$$\mathcal{L}_{L1} = \sum_{h,w,c,t} \left| \hat{\mathbf{B}}_{h,w,c,t} - \mathbf{B}_{h,w,c,t} \right|$$

where $\hat{\mathbf{B}}_{h,w,c,t}$ and $\mathbf{B}_{h,w,c,t}$ are the values of the ground truth background image sequence and the estimated background image sequence on the w-th row, the h-th column, c-th channel, and t-th frame, respectively.

B. Bacteria Segmentation

The second-stage sub-network is a dual-stream 3D encoder-decoder architecture, featuring skip connections from both input streams to improve information transfer for bacteria segmentation. This network takes two sets of inputs: the original OEM image sequence, $\mathbf{I}_m$, and the estimated background image sequence, $\mathbf{B}$. Each sequence is processed in a separate stream. The network initially encodes $\mathbf{I}_m$ and $\mathbf{B}$ into high-level feature representations F_I and F_B, respectively, using lightweight 3D CNN, $\mathbf{\Psi_{Enc}}$. Contrary to the traditional approach of fusing data through concatenation [14], our network adopts a distinct strategy due to the unique characteristics of the input sequences. We calculate the element-wise absolute difference between the two feature maps to produce the feature map F_D. This method emphasizes the change features, indicative of the bacteria. After that cascaded upsampling operations and convolution blocks are applied to F_D to gradually recover the segmentation result $\mathbf{S} \in \mathbb{R}^{t \times C_{out} \times H_0 \times W_0}$. Moreover, similar to the operation at the bottleneck, the absolute difference of the skip-connections from the two streams is used to fuse the encoder features with the decoder features by concatenation for finer segmentation masks with richer spatial details.

The bacteria segmentation sub-network is trained using the same inputs to the background estimation sub-network and their corresponding background estimations. The dice loss $\mathcal{L}_{Dice}$ is utilized for the training.

$$\mathcal{L}_{Dice} = 1 - \frac{2\hat{\mathbf{S}}\mathbf{S} + 1}{\hat{\mathbf{S}} + \mathbf{S} + 1}$$

where $\hat{\mathbf{S}}$ is the true segmentation of the image sequence.

III. Experiment

A. Dataset

In this paper, we conducted experiments using two distinct OEM image sequence datasets. These datasets comprise videos of ex vivo ventilated whole human lungs. The first dataset is made out of lung samples with synthetically generated bacteria proposed by [11] while the other includes lung samples that were infused with bacteria. Both datasets were obtained from distal human lungs, specifically focusing on the alveolar air spaces, using a clinically-ready fiber-based OEM system [15].

The dataset of synthetically generated OEM images featuring bacteria was generated following the approach described by [11]. This technique involves using OEM image sequences without any bacteria as the background. Onto this background, bacteria exhibiting various motion patterns were then added. This process ensures the creation of both the underlying background sequences and precise segmentation masks for each image sequence, making them suitable for training and evaluating supervised machine learning approaches. The dataset is comprised of 1500 sequences, each containing 32 frames with a resolution of 128×128 pixels. The dataset is partitioned into training and testing sets, with 1350 image sequences allocated for training, while the remaining sequences are reserved for evaluating the algorithm's detection performance.

The dataset featuring lungs instilled with bacteria includes a total of 100 frames. Of these, 57 frames capture lung samples before bacteria become visible, while the remaining 43 frames show visible bacteria. Clinicians with expertise in OEM annotated the frames by marking the coordinates of bacteria in the images.

B. Evaluation Metric

Due to the subjective nature of the dataset annotations made by clinicians on lungs with bacteria, we do not consider these annotations as absolute truth. Therefore, for assessing our method, we exclusively use the criterion based on the number of annotations. To compare the number of annotations with the number of detections by our method, we use the Pearson correlation coefficient. This statistical method offers a more objective way to evaluate the performance of our algorithm. Furthermore, we calculate the RMSE to assess the accuracy of our detection counts in relation to the annotations.

Moreover, to quantify the accuracy of the background estimations we utilized the mean absolute error (MAE).

TABLE I
COMPARISON ON REAL BACTERIA DATASET.

Model	Pearson Correlation Coefficient ↑	RMSE ↓
DoG [16]	0.7389	42.7
GSOTH [17]	0.7765	33.4
LoG [18]	0.7929	28.7
Bayesian Approach [10]	0.8242	**4.94**
EmiNet [11]	**0.8922**	7.86
Back2Seg	<u>0.8704</u>	<u>6.81</u>

TABLE II
COMPARISON ON DIFFERENT BACKGROUND ESTIMATION METHODS.

Method	Average MAE ↓
Bayesian Approach [10]	0.0544
3D UNet [20]	0.0393
Back2Seg	**0.0181**

C. Implementation Details

Our model takes as input sequence of size $\mathbf{I}_m \in \mathbb{R}^{t\times C_0\times H_0\times W_0}$, $t = 32$, $C_0 = 1$, and $H_0 = W_0 = 128$. Note that, for simplicity, all notations here have ignored the batch dimension (we use a batch size of 4 for training).

For the background estimation sub-network, $\mathbf{\Phi}_{Enc}$ projects the input samples to feature maps with smaller spatial resolution, $\mathbb{R}^{C_\Phi\times H_\Phi\times W_\Phi}$, with $C_\Phi = 512$, $H_\Phi = \frac{H_0}{16}$, and $W_\Phi = \frac{W_0}{16}$. $\mathbf{\Phi}_{Enc}$ achieves this by stacking 3×3 convolution blocks (3×3 convolutions each followed by a rectified linear unit (ReLu)) with downsampling (2×2 max pooling with strides of two). For the Transformer Encoder, we use multi-head attention with 8 heads and $L = 4$. We train the background estimation sub-network with 100 epochs.

For the bacteria segmentation sub-network, $\mathbf{\Psi}_{Enc}$ encodes the input sequences to spatial resolution, $\mathbb{R}^{t_\Psi\times C_\Psi\times H_\Psi\times W_\Psi}$, with $t_\Psi = \frac{t}{8}$, $C_\Psi = 512$, $H_\Psi = \frac{H_0}{8}$, and $W_\Psi = \frac{W_0}{8}$. $\mathbf{\Psi}_{Enc}$ stacks $3\times 3\times 3$ convolution blocks (similar to $\mathbf{\Phi}_{Enc}$) with downsampling ($2\times 2\times 2$ max pooling with strides of two). Symmetrically, this is then decoded to spatial resolution, $\mathbb{R}^{t\times C_{out}\times H_0\times W_0}$, where $C_{out} = 3$. We train the bacteria segmentation sub-network with 80 epochs.

During training, for both sub-networks, we adopt the Adam optimizer with a learning rate of 0.0001. All experiments were carried out using Pytorch and on a single NVIDIA A100 GPU.

IV. RESULTS AND DISCUSSIONS

A. Comparison with State-of-the-art Methods

The performance of Back2Seg was assessed by comparing its bacteria detection counts with those from manual annotations in image sequences of bacteria-instilled lungs from the real dataset. To quantify Back2Seg detections, the central coordinates of segmented bacteria were identified, each marked as a single detection event. This process enabled the calculation of the total bacteria count per frame. The correlation and RMSE between Back2Seg detection counts and the manually annotated counts were compared against five bacteria detection algorithms: laplacian of Gaussian (LoG) [18], difference of Gaussians (DoG) [16], greyscale opening top-hat filter (GSOTH) [17], [19], Bayesian approach [10], and EmiNet [11]. The comparative results are shown in Table I.

Table I shows that for the proposed approach the correlation coefficient between the counts of trained personnel and the detected number of bacteria is 0.8704, indicating a strong positive correlation. This suggests that the algorithm's detection of bacteria aligns closely with the annotations by trained personnel. EmiNet demonstrates the highest correlation with the annotated counts, but it also has a relatively high RMSE of 7.86 when compared to the Bayesian approach. Although the Bayesian approach yields the lowest RMSE, its correlation with the annotations is not as high as that of EmiNet. Despite not being the top performer in either metric, Back2Seg combines the best aspects of both EmiNet and the Bayesian approach. It improves the correlation with the annotations by 4.62% over the Bayesian approach and reduces the RMSE to 6.81, outperforming EmiNet.

B. Ablation Study

In this section, we conduct ablation studies to evaluate the efficacy of Back2Seg. We examined the impact of the proposed method used for background estimation and the effect of applying the absolute difference during the fusion of skip connections within the segmentation sub-network.

1) Effect of Background Estimation Method: To verify the effect of our proposed background estimation method, we compared it with two different background estimation methods. The first method is the one that is proposed in [10], this method selects a specific video frame and includes two frames before and after it, totaling five frames for analysis. The mean of these frames estimates the background by reducing the visibility of moving bacteria, due to their minimal impact on the overall mean. If adjacent frames are missing, the four closest frames are used instead. This background estimation approach is applied across all video frames. In the second method, we utilized a 3D UNet [20] architecture to encode the input video sequence and decode it to get the estimated background. To evaluate the accuracy of each estimation method, we utilized the synthetic dataset, where the ground truth is known, and calculated the MAE between the estimated background sequence and the ground truth background sequence. Table II shows the comparison between these background estimation methods. It can be seen that our proposed background model demonstrates superior performance compared to other background estimation methods. The method utilized in [10] operates under the assumption that the background remains static, leading to inaccuracies in estimation when there is even minor movement in the scene. In contrast, the 3D UNet approach treats time as an additional spatial dimension and applies 3D convolutions to process the sequence. This method, however, may overlook subtle differences between frames during the decoding process. Our model, on the other hand, encodes each frame independently before employing a Transformer to analyze long-term relationships among the frames. This approach ensures the preservation of essential details in the background estimation process.

TABLE III
COMPARISON ON DIFFERENT FUSION METHODS.

Method	Pearson Correlation Coefficient ↑	RMSE ↓
Concatenation [14]	0.8579	12.21
Absolute Difference	**0.8704**	**6.81**

2) Absolute Difference Fusion: We evaluated our proposed fusion method against the conventional approach, which typically involves fusing data through concatenation [14]. The results, detailed in Table III, show that our approach significantly improves the correlation with, and reduces the RMSE relative to, the annotated values. This improvement occurs because the absolute difference in skip connections explicitly guides the network to compare the disparities between the images. In other words, it aids in detecting significant changes within the image sequences, which primarily involve the bacteria.

V. CONCLUSION

In this paper, we introduce Back2Seg, a novel model for segmenting bacteria in OEM image sequences. Back2Seg uses a dual-stage structure comprising two sub-networks: one dedicated to background estimation, and the other to the segmentation of bacteria. The sub-network for background estimation leverages a hybrid CNN-Transformer model to effectively capture both local and global features. Conversely, the segmentation sub-network is a dual-input network, processing both the original and the background-estimated sequences through a 3D CNN encoder-decoder scheme to accurately segment the bacteria. Notably, this sub-network adopts an absolute difference fusion method over traditional concatenation methods, enhancing segmentation accuracy. Our experiments show that Back2Seg adeptly merges the strengths of both state-of-the-art supervised and unsupervised methods. It notably boosts correlation with annotations by 4.62% compared to unsupervised models and decreases the RMSE by 1.05 surpassing the leading supervised method. Our findings also underscore the effectiveness of the background estimation sub-network and the absolute difference fusion strategy in improving segmentation performance.

ACKNOWLEDGMENT

The human lungs were provided under ethics approved by the independent London - Central Research Ethics Committee 16-LO1883 and NHS Blood and Transplant. We acknowledge the funding of a cross-disciplinary studentship at the University of Edinburgh between the School of Engineering and the College of Medicine and Veterinary Medicine.

This work was also supported by EPSRC through the Healthcare Impact Partnerships, Grant Ref EP/S025987/1, "Next-generation sensing for human in vivo pharmacology-Accelerating drug development in inflammatory diseases".

REFERENCES

[1] T. P. Htun, Y. Sun, H. L. Chua, and J. Pang, "Clinical features for diagnosis of pneumonia among adults in primary care setting: A systematic and meta-review," *Scientific Reports*, vol. 9, no. 1, May 2019.

[2] M. F. Hashmi, S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem, "Efficient pneumonia detection in chest xray images using deep transfer learning," *Diagnostics*, vol. 10, no. 6, p. 417, Jun. 2020.

[3] L. A. Mandell, R. G. Wunderink, A. Anzueto, *et al.*, "Infectious diseases society of america/american thoracic society consensus guidelines on the management of community-acquired pneumonia in adults," *Clinical Infectious Diseases*, vol. 44, no. Supplement_2, S27–S72, Mar. 2007.

[4] J. D. Northrup, R. H. Mach, and M. A. Sellmyer, "Radiochemical approaches to imaging bacterial infections: Intracellular versus extracellular targets," *International Journal of Molecular Sciences*, vol. 20, no. 22, p. 5808, Nov. 2019.

[5] F. S. Fuchs, S. Zirlik, K. Hildner, J. Schubert, M. Vieth, and M. F. Neurath, "Confocal laser endomicroscopy for diagnosing lung cancer in vivo," *European Respiratory Journal*, vol. 41, no. 6, pp. 1401–1408, Sep. 2012.

[6] A. Perperidis, K. Dhaliwal, S. McLaughlin, and T. Vercauteren, "Image computing for fibre-bundle endomicroscopy: A review," *Medical Image Analysis*, vol. 62, p. 101 620, May 2020.

[7] L. Thiberville, S. Moreno-Swirc, T. Vercauteren, E. Peltier, C. Cavé, and G. B. Heckly, "In vivo imaging of the bronchial wall microstructure using fibered confocal fluorescence microscopy," *American Journal of Respiratory and Critical Care Medicine*, vol. 175, no. 1, pp. 22–31, Jan. 2007.

[8] A. K. Eldaly, Y. Altmann, A. Akram, A. Perperidis, K. Dhaliwal, and S. McLaughlin, "Patch-based sparse representation for bacterial detection," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 657–661.

[9] A. K. Eldaly, Y. Altmann, A. Akram, *et al.*, "Bayesian bacterial detection using irregularly sampled optical endomicroscopy images," *Medical Image Analysis*, vol. 57, pp. 18–31, Oct. 2019.

[10] M. Demirel, B. Mills, E. Gaughan, K. Dhaliwal, and J. R. Hopgood, "Bayesian statistical analysis for bacterial detection in pulmonary endomicroscopic fluorescence lifetime imaging," *IEEE Transactions on Image Processing*, vol. 33, pp. 1241–1256, 2024.

[11] M. Demirel, B. Mills, E. Gaughan, K. Dhaliwal, and J. R. Hopgood, "Eminet: Annotation free moving bacteria detection on optical endomicroscopy images," *TechRxiv*, 2024.

[12] H. Lamdouar, W. Xie, and A. Zisserman, "Segmenting invisible moving objects," in *British Machine Vision Conference*, 2021.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2020.

[14] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[15] D. C. Humphries, R. A. O'Connor, H. L. Stewart, *et al.*, "Specific in situ immuno-imaging of pulmonary-resident memory lymphocytes in human lungs," *Frontiers in Immunology*, vol. 14, Feb. 2023.

[16] D. M. Catarious, A. H. Baydush, and C. E. Floyd, "Characterization of difference of gaussian filters in the detection of mammographic regions," *Medical Physics*, vol. 33, no. 11, pp. 4104–4114, Oct. 2006.

[17] Y. Kimori, N. Baba, and N. Morone, "Extended morphological processing: A practical method for automatic spot detection of biological markers from microscopic images," *BMC Bioinformatics*, vol. 11, no. 1, Jul. 2010.

[18] F. He, B. Xiong, C. Sun, and X. Xia, "A laplacian of gaussian-based approach for spot detection in two-dimensional gel electrophoresis images," in *Computer and Computing Technologies in Agriculture IV*, D. Li, Y. Liu, and Y. Chen, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 8–15.

[19] D. Yang, Z. Bai, and J. Zhang, "Infrared weak and small target detection based on top-hat filtering and multi-feature fuzzy decision-making," *Electronics*, vol. 11, no. 21, p. 3549, Oct. 2022.

[20] Ö. undefinediçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, *3d u-net: Learning dense volumetric segmentation from sparse annotation*, 2016.