Efficient Streptococci Segmentation via Divergence-Guided Structured Pruning

Aleksei Samarin, Artem Nazarenko
ITMO University / ISP RAS
St. Petersburg, Russia
avsamarin@itmo.ru, aanazarenko@itmo.ru

Anastasia Mamaeva, Julia Afonina, Aleksei Toropov
ITMO University
St. Petersburg, Russia
asmamaeva@itmo.ru, isafonina@itmo.ru, toropov.ag@hotmail.com

Egor Kotenko
St. Petersburg State University / ISP RAS
St. Petersburg, Russia
kotenkoed@gmail.com

Elena Mikhailova
St. Petersburg State University
St. Petersburg, Russia
e.mikhailova@itmo.ru

Valentin Malykh
International IT University
Almaty, Kazakhstan
valentin.malykh@phystech.edu

Alexander Savelev
St. Petersburg Electrotechnical University / ISP RAS
St. Petersburg, Russia
algsavelev@gmail.com

Alexander Motyko
St. Petersburg Electrotechnical University
St. Petersburg, Russia
aamotyko@etu.ru

***Abstract*—This study addresses efficient semantic segmentation of streptococci in microscopy images of live, unfixed microorganisms. We propose a morphology-aware divergence-guided structured pruning method that combines representation-based component importance with foreground- and multi-scale boundary-sensitive activation information to identify redundant structures while preserving segmentation-relevant spatial features. The method is evaluated on a pixel-level annotated streptococci subset of the previously published MMICD microscopy dataset and compared with random structured, magnitude-based, and divergence-only pruning. At the selected operating point, the proposed approach retains a Dice coefficient of 0.954, an IoU of 0.912, and a Boundary F1 score of 0.936, while reducing model parameters by 50.7%, FLOPs by 47.3%, and inference latency by 31.1%. These results demonstrate that morphology-sensitive component selection improves the accuracy–efficiency trade-off and enables substantial structural compression with only a minor degradation in segmentation quality, making the method suitable for computationally constrained microscopy applications.**

I. Introduction

Streptococci are a heterogeneous group of Gram-positive coccoid bacteria that occupy diverse ecological niches and include both commensal and clinically significant microorganisms. Oral commensal streptococci contribute to the organization and stability of microbial communities, whereas other representatives of the genus are associated with infectious diseases. Streptococci are also important in food biotechnology; for example, *Streptococcus thermophilus* is widely used in dairy fermentation [1], [2]. Reliable detection and quantitative characterization of these microorganisms are therefore relevant to biomedical research, microbiological monitoring, and food-production analysis.

From a microscopy perspective, streptococci are characterized by small coccoid or approximately ovoid cells that frequently form pairs and chain-like arrangements [3]. Their morphology creates a challenging segmentation problem: target structures may occupy only a small fraction of the image, neighboring cocci can form visually connected regions, and foreground contours may become weak or partially obscured. These difficulties are amplified in live, unfixed microscopy by motion blur, low contrast, non-uniform illumination, optical noise, and heterogeneous background structures. Reliable analysis therefore requires accurate pixel-level localization of streptococcal regions and preservation of their fine morphological contours.

Semantic segmentation provides the spatial information required for estimating microorganism area, morphology, density, and spatial organization. Modern deep-learning methods have substantially improved bacterial image segmentation and can accommodate greater variability in morphology and imaging conditions than conventional thresholding and morphology-based pipelines [4]–[6]. Nevertheless, accurate segmentation of compact and boundary-dominated bacterial structures still depends on preserving fine local features while maintaining sufficient contextual information.

Context-aware segmentation architectures address this requirement through attention, non-local operations, and transformer-based mechanisms that model dependencies between spatially distant features [7], [8]. Their increased representational capacity, however, is typically accompanied by higher parameter counts, computational complexity, memory consumption, and inference latency. This limits their practical use in high-throughput microscopy and resource-constrained laboratory environments. Consequently, an effective segmentation model should provide not only high pixel-level accuracy but also a favorable balance between prediction quality and

computational cost.

Structured neural network pruning reduces model complexity by removing complete filters, channels, or computational blocks [9], [10]. The central difficulty is determining which components can be removed without eliminating spatially important information. Conventional pruning criteria are commonly based on parameter magnitude, activation statistics, or globally aggregated importance estimates. For microscopy segmentation, such criteria can underestimate channels whose overall contribution is limited but whose activations are concentrated around small foreground regions or narrow object contours. Compression based only on global redundancy may therefore disproportionately affect features that are important for preserving streptococcal morphology.

The objective of this study is to reduce the computational complexity of streptococci segmentation models while preserving morphology-sensitive representations and pixel-level prediction quality. We propose a morphology-aware divergence-guided structured pruning method that combines representation-based component importance with foreground- and multi-scale boundary-sensitive activation statistics. Low-importance components are removed under structural and validation constraints, followed by representation-preserving fine-tuning. The method is evaluated on a pixel-level annotated streptococci subset derived from the previously published MMICD microscopy dataset [11]. Segmentation accuracy, boundary preservation, parameter reduction, FLOPs, and inference latency are jointly analyzed to determine whether morphology-sensitive information improves the accuracy–efficiency trade-off of structured pruning.

II. Related Work

Existing approaches relevant to streptococci segmentation can be broadly grouped into conventional image-processing methods, convolutional encoder–decoder networks, specialized cell and bacterial segmentation models, context-aware architectures, and model-compression techniques. These directions address complementary aspects of microscopy image analysis; however, morphology-sensitive bacterial segmentation and computationally efficient model compression have largely been studied separately.

Classical microscopy segmentation methods rely on intensity thresholding, edge detection, region growing, watershed transformation, connected-component analysis, and mathematical morphology. Their main advantages are low computational cost and direct interpretability. Such approaches remain effective when microorganisms exhibit sufficiently high contrast and are well separated from the background. However, their performance is strongly dependent on manually selected thresholds and assumptions about object intensity, scale, and contour structure. Consequently, they are sensitive to illumination variation, blur, weak contrast, and heterogeneous background regions. These limitations become particularly important in live, unfixed microscopy, where acquisition artifacts and microorganism motion reduce the reliability of fixed image-processing rules.

Convolutional encoder–decoder networks substantially improved biomedical image segmentation by learning task-specific representations directly from annotated data. U-Net combines progressively abstract encoder features with higher-resolution information propagated through skip connections [4], whereas U-Net++ introduces nested skip pathways to improve interaction between encoder and decoder representations [12]. DeepLabV3+ incorporates atrous convolutions and multi-scale context aggregation to capture structures at different spatial scales [13]. Although these architectures are more flexible than conventional image-processing pipelines, repeated spatial downsampling can suppress information associated with small foreground regions and fine contours. Increasing network depth or introducing additional multi-scale processing can mitigate this effect, but generally increases computational complexity and does not explicitly account for microorganism-specific morphology.

Several methods have been developed specifically for cellular and bacterial microscopy. Cellpose introduced a generalist deep-learning framework for cell segmentation across heterogeneous microscopy data [14]. DeepBacs demonstrated deep-learning pipelines for bacterial image analysis, including segmentation of bright-field and fluorescence microscopy images [5]. Omnipose further addressed morphology-dependent segmentation by introducing representations designed to handle cells with diverse shapes and geometries [6]. Recent studies have also explored the transfer of foundation and vision–language models to bacterial microscopy. Domain adaptation of SAM-based models has been investigated for monococci segmentation under microscopy-specific appearance variability [15], while Grounding DINO has been applied to bacilli localization using microscopy-oriented input adaptation [16]. These studies demonstrate the potential of large pre-trained vision models for microbial image analysis, but primarily address prediction robustness and domain adaptation rather than morphology-aware structural compression.

A complementary direction focuses on increasing contextual information available to the segmentation model. Non-local operations explicitly model dependencies between spatially distant feature locations [7], while transformer-based architectures use self-attention to capture broader spatial relationships [8], [17]. Such contextual information can be useful in microscopy scenes containing compact, spatially clustered, or chain-like bacterial structures. Our previous work on diplococci segmentation also investigated modified non-local blocks for microorganism-specific contextual representation [18]. The main limitation of these approaches is their computational overhead: attention operations, projection layers, and high-dimensional contextual representations can substantially increase FLOPs, memory consumption, and inference latency. This creates a trade-off between representational capacity and practical computational efficiency.

Model compression provides a complementary mechanism for addressing this trade-off. Conventional structured pruning removes complete filters, channels, or other computational components according to criteria such as parameter magnitude,

activation strength, sensitivity, or estimated loss variation [9], [10], [19]. More recent studies have applied structured pruning directly to segmentation networks. López-González et al. investigated filter- and layer-pruning strategies for architectures including U-Net, DeepLabV3+, and SegNet using feature-map-based importance estimation [20]. Sauron U-Net reduces redundant feature maps during training by encouraging feature similarity and determining layer-dependent pruning thresholds [21]. These studies demonstrate that segmentation networks contain substantial structural redundancy, but their pruning criteria are primarily based on generic redundancy or feature statistics rather than microorganism-specific morphological information.

An important distinction therefore arises between classification-oriented and segmentation-oriented pruning. Wu et al. showed that pruning criteria developed for image recognition can be suboptimal for semantic segmentation because dense prediction is intrinsically location-sensitive. Their SIRFP approach estimates spatial redundancy from non-pooled feature maps and preserves spatial information during filter selection [22]. This represents an important step toward segmentation-aware compression. However, spatial awareness alone does not explicitly distinguish between background redundancy and features concentrated around biologically relevant foreground regions or object contours. For bacterial microscopy, this distinction is important because globally weak activations may still encode information associated with small or boundary-dominated structures.

Our previous studies address complementary parts of this problem. Streptococci recognition was investigated using interpretable taxonomy-based visual characteristics [3], while MMICD extended microorganism recognition to a multi-class microscopy setting [11]. IDAP++ subsequently introduced divergence-aware structural compression based on changes in internal information flow and joint component- and layer-level optimization [23]. These studies provide, respectively, a morphology-oriented representation perspective and a divergence-based approach to structural importance estimation, but they do not address their integration within pixel-level streptococci segmentation.

The resulting research gap lies at the intersection of morphology-sensitive bacterial segmentation and segmentation-aware model compression. Existing bacterial segmentation methods predominantly focus on predictive quality, whereas structured pruning methods primarily estimate redundancy or generic component importance. The present study connects these directions by combining representation divergence with foreground- and multi-scale boundary-sensitive activation information. The resulting pruning criterion is designed to reduce computational redundancy while preferentially preserving representations associated with compact and boundary-dominated streptococcal regions.

III. Problem Statement

Let

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \qquad x_i \in \mathbb{R}^{H\times W\times C}, \qquad y_i \in \{0,1\}^{H\times W}, \tag{1}$$

denote a microscopy segmentation dataset, where $y_i(p) = 1$ indicates a streptococcal foreground pixel and $y_i(p) = 0$ denotes background.

A segmentation network is represented by

$$f_\theta : \mathbb{R}^{H\times W\times C} \to [0,1]^{H\times W}, \tag{2}$$

and produces the probability map

$$\hat{y}_i = f_\theta(x_i). \tag{3}$$

The binary prediction is obtained as

$$\tilde{y}_i(p) = \mathbb{I}\left[\hat{y}_i(p) \geq \tau\right]. \tag{4}$$

The baseline network is trained by minimizing

$$\theta^* = \arg\min_\theta \frac{1}{|\mathcal{D}_{train}|} \sum_{(x_i,y_i)\in\mathcal{D}_{train}} \mathcal{L}_{seg}\left(f_\theta(x_i), y_i\right), \tag{5}$$

where the segmentation loss combines binary cross-entropy and Dice loss:

$$\mathcal{L}_{seg} = \lambda_{BCE}\mathcal{L}_{BCE} + \lambda_{Dice}\mathcal{L}_{Dice}. \tag{6}$$

The individual terms are defined as

$$\mathcal{L}_{BCE} = -\frac{1}{|\Omega|}\sum_{p\in\Omega}\left[y(p)\log\hat{y}(p) + (1-y(p))\log(1-\hat{y}(p))\right], \tag{7}$$

and

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{p\in\Omega} y(p)\hat{y}(p) + \delta}{\sum_{p\in\Omega} y(p) + \sum_{p\in\Omega}\hat{y}(p) + \delta}, \tag{8}$$

where $\delta > 0$ ensures numerical stability.

Segmentation quality is evaluated using

$$Dice = \frac{2TP}{2TP + FP + FN}, \qquad IoU = \frac{TP}{TP + FP + FN}. \tag{9}$$

The objective of this study is not only to preserve segmentation quality but also to reduce the structural and computational complexity of the trained network. Following the structured-pruning paradigm [9], [10], let

$$\mathcal{G} = \{g_1, g_2, \ldots, g_K\} \tag{10}$$

denote the removable network components. Depending on the architecture, these may correspond to convolutional channels, filters, attention components, or contextual blocks.

A pruning configuration is represented by

$$m = (m_1, \ldots, m_K), \qquad m_k \in \{0,1\}, \tag{11}$$

where $m_k = 1$ retains component g_k and $m_k = 0$ removes it. The resulting compressed model is denoted by $f_{\theta',m}$.

For microscopy segmentation, globally small representation changes may still occur in spatially important foreground or

boundary regions. We therefore additionally constrain distortion of intermediate representations. Let $T_l(x)$ and $T_l^m(x)$ denote the features produced by the original and compressed models at stage l. Their normalized representation distortion is defined as

$$\Delta_{\mathcal{F}} = \frac{1}{|\mathcal{S}|} \sum_{l \in \mathcal{S}} \mathbb{E}_{x \in \mathcal{D}_{val}} \left[\frac{\|\psi_l(T_l(x)) - \psi_l(T_l^m(x))\|_2}{\|\psi_l(T_l(x))\|_2 + \eta} \right], \tag{12}$$

where $\mathcal{S}$ is the set of monitored stages, $\psi_l(\cdot)$ aligns representation dimensions when necessary, and $\eta > 0$ prevents numerical instability.

The compression problem is finally formulated as

$$(m^*, \theta'^*) = \arg\min_{m,\theta'} \mathcal{C}\left(f_{\theta',m}\right) \tag{13}$$

subject to

$$\Delta Q \le \varepsilon, \qquad \Delta_{\mathcal{F}} \le \gamma, \qquad m \in \mathcal{M}_{struct}, \tag{14}$$

where

$$\Delta Q = Q\left(f_{\theta^*}, \mathcal{D}_{val}\right) - Q\left(f_{\theta',m}, \mathcal{D}_{val}\right). \tag{15}$$

Here, $\mathcal{C}$ denotes the computational cost of the compressed model, characterized experimentally by parameter count, FLOPs, and inference latency; ε is the maximum admissible segmentation-quality degradation, γ bounds representation distortion, and $\mathcal{M}_{struct}$ contains structurally valid pruning configurations.

Thus, the task is to obtain a structurally compressed segmentation model that reduces computational cost while preserving both pixel-level prediction quality and representations associated with streptococcal foreground regions and their contours.

IV. Proposed Solution

The proposed method compresses a trained streptococci segmentation network by identifying structural components that have limited influence on internal representations while preserving features concentrated within foreground regions and along their contours. The method combines representation-based importance, morphology-sensitive activation statistics, and a weak magnitude prior, followed by structured pruning and representation-preserving fine-tuning. Fig. 1 shows an overview of the proposed framework.

A. Representation-Based Importance

Let the trained baseline network be represented as

$$f_{\theta^*}(x) = F_L \circ F_{L-1} \circ \ldots \circ F_1(x), \tag{16}$$

with intermediate representations

$$T_l(x) = F_l(T_{l-1}(x)), \qquad T_0(x) = x. \tag{17}$$

For a removable structural component $g_{l,k}$, its representation-based importance is estimated by temporarily suppressing the component while keeping the remaining network unchanged:

$$D_{l,k} = \mathbb{E}_{x \in \mathcal{D}_{val}} \left[\frac{\left\| \psi_l(T_l(x)) - \psi_l(T_l^{(-k)}(x)) \right\|_2}{\|\psi_l(T_l(x))\|_2 + \eta} \right], \tag{18}$$

where $T_l^{(-k)}(x)$ denotes the representation obtained when component $g_{l,k}$ is disabled and $\eta > 0$ ensures numerical stability. A larger $D_{l,k}$ indicates a stronger influence of the component on the internal representation and therefore a higher importance for preservation. This representation-perturbation principle follows the divergence-aware compression perspective introduced in our previous work [23].

B. Morphology-Aware Importance

Representation change alone does not indicate whether the affected information is associated with homogeneous background regions or segmentation-relevant structures. We therefore introduce a morphology-sensitive term that emphasizes both streptococcal foreground regions and their contours.

For binary mask y, a boundary map at scale r is defined as

$$b_r(y) = \mathrm{Dilate}_r(y) - \mathrm{Erode}_r(y), \tag{19}$$

and the multi-scale boundary representation is

$$b_{ms}(y) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} b_r(y), \qquad \mathcal{R} = \{1, 2, 3\}. \tag{20}$$

The resulting spatial weighting map is

$$\omega(y) = 1 + \beta y + \kappa b_{ms}(y), \tag{21}$$

where β controls foreground weighting and κ determines the additional contribution of boundary regions.

For the activation associated with component $g_{l,k}$, denoted by $A_{l,k}(x)$, morphology-aware importance is defined as

$$M_{l,k} = \mathbb{E}_{(x,y) \in \mathcal{D}_{val}} \left[\frac{\|\omega_l(y) \odot A_{l,k}(x)\|_F}{\|\omega_l(y)\|_F + \eta} \right], \tag{22}$$

where $\omega_l(y)$ is resized to match the spatial resolution of the corresponding activation map. This term increases the importance of components whose responses are concentrated within small foreground structures or near their contours, even when their global activation magnitude is relatively low.

C. Combined Pruning Criterion

A magnitude-based score is additionally used as a weak structural prior:

$$W_{l,k}^{mag} = \|W_{l,k}\|_1, \tag{23}$$

where $W_{l,k}$ denotes the parameter tensor associated with component $g_{l,k}$.

Because the three criteria have different numerical ranges, they are normalized independently within each prunable layer:

$$\bar{z}_{l,k} = \frac{z_{l,k} - \min_j z_{l,j}}{\max_j z_{l,j} - \min_j z_{l,j} + \eta}. \tag{24}$$

The final component importance is

$$S_{l,k} = \lambda_D \bar{D}_{l,k} + \lambda_M \bar{M}_{l,k} + \lambda_W \bar{W}_{l,k}, \qquad \lambda_D + \lambda_M + \lambda_W = 1. \tag{25}$$

In the proposed configuration,

$$\lambda_D = 0.55, \qquad \lambda_M = 0.35, \qquad \lambda_W = 0.10. \tag{26}$$

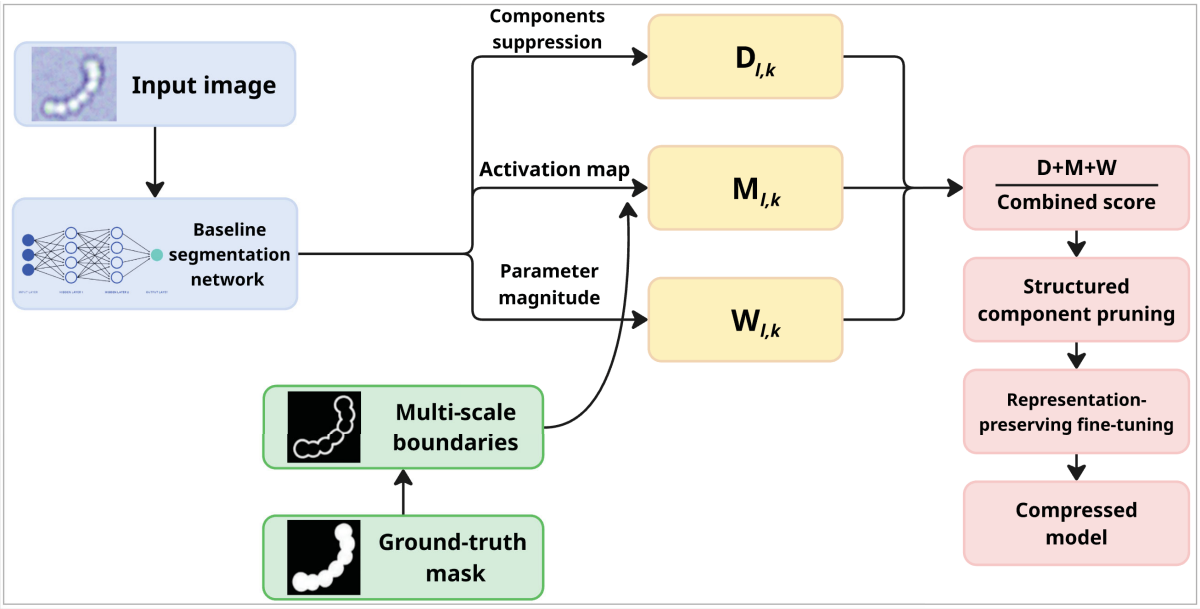


Fig. 1. Overview of the proposed morphology-aware divergence-guided structured pruning framework

For target pruning ratio p_l, the layer-specific threshold is

$$\tau_l = \text{Percentile}\left(\{S_{l,k}\}_{k=1}^{K_l}, p_l\right), \tag{27}$$

and the corresponding structural mask is

$$m_{l,k} = \mathbb{I}\left[S_{l,k} \geq \tau_l\right]. \tag{28}$$

Components below the threshold are removed together with structurally dependent dimensions. Skip and residual connections are adjusted consistently to preserve architectural validity. To prevent degenerate layer configurations, at least 25% of the original structural units are retained in each prunable layer.

D. Contextual Block Pruning

After component-level pruning, higher-level contextual modules are evaluated separately when supported by the architecture. For a removable block B_l, its contribution is estimated as

$$S_l^{block} = \mathbb{E}_{x \in \mathcal{D}_{val}}\left[\frac{\|\psi_l(B_l(T_l(x))) - \psi_l(T_l(x))\|_2}{\|\psi_l(T_l(x))\|_2 + \eta}\right]. \tag{29}$$

Blocks with low contribution are considered candidates for removal. A removed block is replaced by an identity mapping whenever dimensions are preserved, or by a lightweight 1×1 projection when dimensional adaptation is required. Block removal is accepted only when architectural consistency is maintained and the validation-quality constraint remains satisfied.

E. Representation-Preserving Fine-Tuning

After structural pruning, the compressed network is fine-tuned using

$$\mathcal{L}_{FT} = \mathcal{L}_{seg} + \lambda_R \mathcal{L}_{rep}, \tag{30}$$

where

$$\mathcal{L}_{rep} = \frac{1}{|\mathcal{S}|}\sum_{l \in \mathcal{S}} \mathbb{E}_{x \in \mathcal{D}_{train}}\left[\frac{\|\psi_l(T_l(x)) - \psi_l(T_l^m(x))\|_2^2}{\|\psi_l(T_l(x))\|_2^2 + \eta}\right]. \tag{31}$$

The original uncompressed network remains frozen and serves as a representation reference during fine-tuning. A pruning configuration is accepted only if

$$\Delta Q_{val} = Q(f_{\theta^*}, \mathcal{D}_{val}) - Q(f_{\theta',m}, \mathcal{D}_{val}) \leq \varepsilon. \tag{32}$$

If this condition is violated, a less aggressive pruning ratio is applied.

Thus, the proposed method combines representation perturbation with morphology-sensitive spatial information when ranking removable network components. This allows structural redundancy to be reduced while preferentially preserving representations associated with streptococcal foreground regions and their contours.

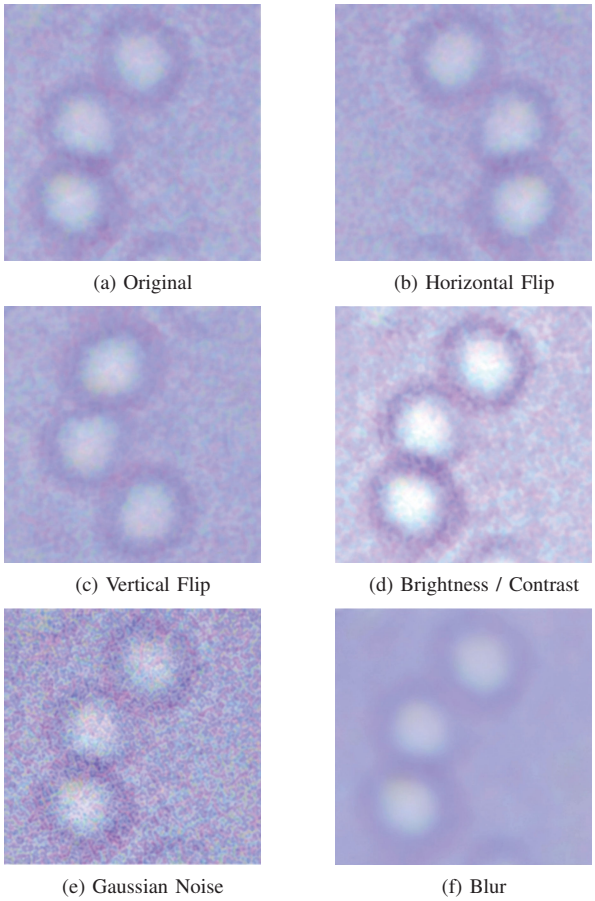
V. Experimental Setup and Results

A. Dataset

The experiments are conducted using the streptococci class of the previously published MMICD microscopy dataset [11]. The selected subset contains 3808 microscopy images of streptococci acquired from live, unfixed samples using a Levenhuk MED D30T microscope.

Representative examples of the original microscopy image and a subset of the data augmentation operations used in the training subset are shown in Fig. 2.

For this semantic segmentation task, images are associated with binary pixel-level masks, where streptococcal regions are treated as foreground and all remaining pixels as background. The dataset is divided into training, validation, and test subsets



(a) Original (b) Horizontal Flip (c) Vertical Flip (d) Brightness / Contrast (e) Gaussian Noise (f) Blur

Fig. 2. Examples of microscopy images from the streptococci subset and a subset of the augmentations used in training

using a 5:1:1 ratio, resulting in 2720 training images, 544 validation images, and 544 test images. The split is performed at the source-image level before augmentation to prevent information leakage between the subsets.

All images and corresponding masks are resized to 256×256 pixels. Data augmentation is applied exclusively to the training subset and includes random horizontal and vertical flipping, rotation, brightness and contrast variation, Gaussian noise, and blur. No augmentation is applied to the validation or test subsets.

B. Experimental Protocol

The baseline evaluation covers convolutional encoder–decoder, transformer-based, hybrid, and promptable segmentation architectures. We consider U-Net and U-Net++ [4], [12], their ResNet- and EfficientNet-based variants [24], [25], DeepLabV3+ [13], SETR [26], SegFormer [27], Mask2Former [28], TransUNet [8], Swin-Unet [29], Dual-Stream CoAtNet [30], and an attention-enhanced U-Net. SAM [31], SAM2 [32], and MedSAM [33] are additionally evaluated as promptable segmentation baselines.

Structured compression experiments compare random structured pruning, magnitude-based pruning, divergence-only pruning, and the proposed morphology-aware divergence-guided strategy. Pruning is evaluated across compatible trainable architectures. Promptable foundation models are excluded from the pruning comparison unless they are used in a fully trainable and structurally compatible configuration. The attention-enhanced U-Net is used as the main architecture for the detailed ablation and compression-level analyses.

All trainable models are optimized using AdamW with an initial learning rate of 10^{-4} and a batch size 16. The training objective combines binary cross-entropy and Dice loss. The morphology-aware weighting parameters are fixed to $\beta = 1$ and $\kappa = 2$, the representation-preservation coefficient is $\lambda_R = 0.1$, and the maximum admissible validation Dice degradation is $\varepsilon = 0.02$. The pruning-score weights are fixed to $\lambda_D = 0.55$, $\lambda_M = 0.35$, and $\lambda_W = 0.10$. Target pruning ratios are 10%, 25%, 50%, and 75%.

Segmentation quality is evaluated using IoU, Dice coefficient, and Boundary F1 score. Computational efficiency is characterized by parameter count, FLOPs, and inference time. Inference time is measured using batch size 1 and 256×256 input images on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU. All measurements are performed in FP32 precision using the same PyTorch and CUDA environment. For each model, 100 warm-up iterations are followed by 1000 timed forward passes, with CUDA synchronization before and after each measurement.

C. Results

The evaluation proceeds in three stages: uncompressed baselines are compared to select a backbone, an ablation study isolates the contribution of each importance criterion, and increasing pruning strength is examined to select an operating point for comparison against alternative pruning strategies.

Table I compares segmentation quality and computational cost of the uncompressed models. The Attention-enhanced U-Net attains the highest Dice (0.958) and IoU (0.920) at a lower computational cost than transformer-based models such as SETR and TransUNet, while foundation models (SAM, SAM2, MedSAM) trail despite their larger size, reflecting a domain gap with low-contrast microscopy images. It is therefore adopted as the backbone for the pruning experiments.

Table II isolates the contributions of magnitude, divergence-based, and morphology-aware importance. Magnitude alone yields the lowest accuracy; adding either divergence or morphology individually improves all metrics, and combining them raises BF1 from 0.911 to 0.934. The full method, which adds the weak magnitude prior, achieves the best overall trade-off, indicating the three signals are complementary.

Table III evaluates increasing pruning strength. Quality is essentially unaffected up to 25% pruning, after which accuracy degrades progressively; BF1 drops more sharply than IoU/Dice, from 0.946 to 0.903 at 75% pruning, showing boundary information degrades first under aggressive compression and motivating a moderate operating point.

Table IV and Table V compare pruning strategies across trainable architectures at matched compression ratios.

TABLE I. PERFORMANCE OF UNCOMPRESSED SEGMENTATION MODELS ON THE STREPTOCOCCI TEST SUBSET

Model	Params, M	FLOPs, G	Time, ms	IoU	Dice
U-Net	31.0	54.7	3.8	0.884 ± 0.006	0.938 ± 0.004
U-Net++	36.6	68.9	4.9	0.891 ± 0.005	0.942 ± 0.003
U-Net ResNet-18	14.3	28.1	3.2	0.887 ± 0.006	0.940 ± 0.004
U-Net ResNet-50	32.6	46.8	4.4	0.898 ± 0.005	0.946 ± 0.003
U-Net EfficientNet-B4	19.5	23.9	4.7	0.902 ± 0.005	0.948 ± 0.003
U-Net++ EfficientNet-B4	23.9	31.6	5.8	0.906 ± 0.004	0.951 ± 0.003
DeepLabV3+	40.4	38.7	4.6	0.892 ± 0.006	0.943 ± 0.004
SETR	86.2	57.3	7.8	0.889 ± 0.007	0.941 ± 0.004
SegFormer	27.4	17.1	3.9	0.904 ± 0.004	0.950 ± 0.003
Mask2Former	44.1	72.8	8.4	0.910 ± 0.004	0.953 ± 0.002
TransUNet	105.3	48.5	6.9	0.908 ± 0.005	0.952 ± 0.003
Swin-Unet	27.2	25.4	5.7	0.912 ± 0.004	0.954 ± 0.002
Dual-Stream CoAtNet	48.7	36.9	6.2	0.916 ± 0.004	0.956 ± 0.002
SAM	641.1	2973.4	89.6	0.872 ± 0.008	0.932 ± 0.005
SAM2	224.4	712.6	34.8	0.889 ± 0.007	0.941 ± 0.004
MedSAM	93.7	168.3	18.6	0.900 ± 0.005	0.947 ± 0.003
Attention-enhanced U-Net	34.9	59.6	4.5	0.920 ± 0.003	0.958 ± 0.002

TABLE II. ABLATION STUDY OF THE PROPOSED PRUNING CRITERION

Variant	Params red., %	FLOPs red., %	Time, ms	IoU	Dice	BF1
No pruning	0	0	4.5	0.920 ± 0.005	0.958 ± 0.002	0.946 ± 0.003
Magnitude only	49.1	46.0	3.3	0.899 ± 0.005	0.947 ± 0.003	0.911 ± 0.006
Divergence only	49.4	46.6	3.2	0.907 ± 0.004	0.951 ± 0.003	0.925 ± 0.005
Morphology only	48.8	45.9	3.3	0.905 ± 0.005	0.950 ± 0.003	0.929 ± 0.005
Divergence + morphology	49.7	46.8	3.2	0.910 ± 0.004	0.953 ± 0.002	0.934 ± 0.004
Full proposed method	50.7	47.3	3.1	0.912 ± 0.004	0.954 ± 0.002	0.936 ± 0.004

TABLE III. COMPRESSION-LEVEL ANALYSIS OF THE PROPOSED METHOD

Pruning	Params, M	FLOPs, G	IoU	Dice	BF1
0%	34.9	59.6	0.920 ± 0.003	0.958 ± 0.002	0.946 ± 0.003
10%	30.1	53.2	0.920 ± 0.003	0.958 ± 0.002	0.945 ± 0.003
25%	25.3	46.1	0.918 ± 0.004	0.957 ± 0.002	0.942 ± 0.004
50%	17.2	31.4	0.912 ± 0.004	0.954 ± 0.002	0.936 ± 0.004
75%	8.9	17.0	0.889 ± 0.006	0.941 ± 0.004	0.903 ± 0.007

TABLE IV. COMPARISON OF STRUCTURED PRUNING STRATEGIES ACROSS TRAINABLE SEGMENTATION MODELS ON THE STREPTOCOCCI TEST SUBSET (PART 1)

Pruning method	Params red., %	FLOPs red., %	Time, ms	IoU	Dice	BF1
U-Net						
None	0	0	3.8	0.884 ± 0.006	0.938 ± 0.004	0.917 ± 0.005
Random structured	49.6	46.8	2.8	0.850 ± 0.009	0.919 ± 0.006	0.867 ± 0.010
Magnitude-based	49.2	46.3	2.8	0.866 ± 0.007	0.928 ± 0.005	0.889 ± 0.008
Divergence-only	49.8	47.0	2.7	0.874 ± 0.006	0.933 ± 0.004	0.902 ± 0.006
Morphology-aware proposed	50.3	47.5	2.7	0.878 ± 0.005	0.935 ± 0.004	0.910 ± 0.005
U-Net++						
None	0	0	4.9	0.891 ± 0.005	0.942 ± 0.003	0.923 ± 0.005
Random structured	48.8	44.9	3.8	0.856 ± 0.008	0.922 ± 0.006	0.873 ± 0.009
Magnitude-based	49.1	45.5	3.7	0.873 ± 0.007	0.932 ± 0.004	0.897 ± 0.007
Divergence-only	49.6	46.2	3.6	0.881 ± 0.006	0.937 ± 0.004	0.909 ± 0.006
Morphology-aware proposed	50.1	46.8	3.6	0.886 ± 0.005	0.940 ± 0.003	0.917 ± 0.005
U-Net ResNet-18						
None	0	0	3.2	0.887 ± 0.006	0.940 ± 0.004	0.920 ± 0.005
Random structured	45.7	42.6	2.5	0.852 ± 0.009	0.920 ± 0.006	0.869 ± 0.010
Magnitude-based	46.2	43.1	2.5	0.869 ± 0.007	0.930 ± 0.005	0.892 ± 0.008
Divergence-only	46.8	43.9	2.4	0.878 ± 0.006	0.935 ± 0.004	0.906 ± 0.006
Morphology-aware proposed	47.3	44.5	2.4	0.882 ± 0.005	0.937 ± 0.004	0.914 ± 0.005

TABLE V. COMPARISON OF STRUCTURED PRUNING STRATEGIES ACROSS TRAINABLE SEGMENTATION MODELS ON THE STREPTOCOCCI TEST SUBSET (PART 2)

Pruning method	**Params red., %**	**FLOPs red., %**	**Time, ms**	**IoU**	**Dice**	**BF1**
U-Net ResNet-50						
None	0	0	4.4	0.898 ± 0.005	0.946 ± 0.003	0.929 ± 0.004
Random structured	46.1	43.8	3.3	0.864 ± 0.008	0.927 ± 0.005	0.880 ± 0.009
Magnitude-based	46.7	44.3	3.3	0.881 ± 0.006	0.937 ± 0.004	0.905 ± 0.007
Divergence-only	47.2	45.0	3.2	0.889 ± 0.005	0.941 ± 0.003	0.917 ± 0.005
Morphology-aware proposed	47.8	45.7	3.2	0.894 ± 0.004	0.944 ± 0.003	0.924 ± 0.004
U-Net EfficientNet-B4						
None	0	0	4.7	0.902 ± 0.005	0.948 ± 0.003	0.932 ± 0.004
Random structured	42.9	39.8	3.8	0.869 ± 0.008	0.930 ± 0.005	0.886 ± 0.009
Magnitude-based	43.5	40.4	3.7	0.885 ± 0.006	0.939 ± 0.004	0.908 ± 0.007
Divergence-only	44.0	41.1	3.7	0.893 ± 0.005	0.943 ± 0.003	0.920 ± 0.005
Morphology-aware proposed	44.6	41.8	3.6	0.898 ± 0.004	0.946 ± 0.003	0.927 ± 0.004
U-Net++ EfficientNet-B4						
None	0	0	5.8	0.906 ± 0.004	0.951 ± 0.003	0.936 ± 0.004
Random structured	43.7	40.9	4.6	0.873 ± 0.008	0.932 ± 0.005	0.890 ± 0.009
Magnitude-based	44.1	41.5	4.5	0.889 ± 0.006	0.941 ± 0.004	0.913 ± 0.006
Divergence-only	44.8	42.2	4.4	0.898 ± 0.005	0.946 ± 0.003	0.925 ± 0.005
Morphology-aware proposed	45.3	42.9	4.4	0.902 ± 0.004	0.948 ± 0.003	0.932 ± 0.004
DeepLabV3+						
None	0	0	4.6	0.892 ± 0.006	0.943 ± 0.004	0.925 ± 0.005
Random structured	47.2	45.0	3.4	0.857 ± 0.009	0.923 ± 0.006	0.875 ± 0.010
Magnitude-based	47.7	45.6	3.3	0.875 ± 0.007	0.933 ± 0.005	0.900 ± 0.008
Divergence-only	48.3	46.2	3.3	0.883 ± 0.006	0.938 ± 0.004	0.913 ± 0.006
Morphology-aware proposed	48.8	46.9	3.2	0.888 ± 0.005	0.941 ± 0.003	0.920 ± 0.005
SETR						
None	0	0	7.8	0.889 ± 0.007	0.941 ± 0.004	0.921 ± 0.006
Random structured	41.5	38.2	6.2	0.850 ± 0.010	0.919 ± 0.007	0.865 ± 0.011
Magnitude-based	42.1	38.9	6.1	0.868 ± 0.008	0.929 ± 0.005	0.891 ± 0.009
Divergence-only	42.8	39.7	6.0	0.879 ± 0.006	0.936 ± 0.004	0.907 ± 0.007
Morphology-aware proposed	43.4	40.5	5.9	0.884 ± 0.005	0.938 ± 0.004	0.915 ± 0.006
SegFormer						
None	0	0	3.9	0.904 ± 0.004	0.950 ± 0.003	0.934 ± 0.004
Random structured	44.0	40.6	3.1	0.871 ± 0.008	0.931 ± 0.005	0.888 ± 0.009
Magnitude-based	44.6	41.3	3.0	0.887 ± 0.006	0.940 ± 0.004	0.911 ± 0.007
Divergence-only	45.1	42.0	3.0	0.896 ± 0.005	0.945 ± 0.003	0.924 ± 0.005
Morphology-aware proposed	45.7	42.7	2.9	0.900 ± 0.004	0.947 ± 0.003	0.930 ± 0.004
Mask2Former						
None	0	0	8.4	0.910 ± 0.004	0.953 ± 0.002	0.940 ± 0.004
Random structured	40.8	37.6	6.9	0.876 ± 0.008	0.934 ± 0.005	0.893 ± 0.009
Magnitude-based	41.4	38.2	6.8	0.893 ± 0.006	0.943 ± 0.004	0.916 ± 0.006
Divergence-only	42.0	39.0	6.7	0.901 ± 0.005	0.948 ± 0.003	0.929 ± 0.005
Morphology-aware proposed	42.6	39.7	6.6	0.906 ± 0.004	0.951 ± 0.003	0.936 ± 0.004
TransUNet						
None	0	0	6.9	0.908 ± 0.005	0.952 ± 0.003	0.938 ± 0.004
Random structured	43.3	39.9	5.5	0.872 ± 0.008	0.932 ± 0.005	0.888 ± 0.009
Magnitude-based	43.9	40.6	5.4	0.890 ± 0.006	0.942 ± 0.004	0.913 ± 0.007
Divergence-only	44.5	41.4	5.3	0.899 ± 0.005	0.947 ± 0.003	0.926 ± 0.005
Morphology-aware proposed	45.0	42.0	5.2	0.904 ± 0.004	0.950 ± 0.003	0.934 ± 0.004
Swin-Unet						
None	0	0	5.7	0.912 ± 0.004	0.954 ± 0.002	0.941 ± 0.004
Random structured	42.6	39.7	4.6	0.878 ± 0.008	0.935 ± 0.005	0.894 ± 0.009
Magnitude-based	43.2	40.4	4.5	0.895 ± 0.006	0.945 ± 0.004	0.918 ± 0.006
Divergence-only	43.8	41.1	4.4	0.904 ± 0.005	0.950 ± 0.003	0.931 ± 0.005
Morphology-aware proposed	44.4	41.9	4.3	0.908 ± 0.004	0.952 ± 0.003	0.938 ± 0.004
Dual-Stream CoAtNet						
None	0	0	6.2	0.916 ± 0.004	0.956 ± 0.002	0.944 ± 0.003
Random structured	43.8	40.5	5.0	0.881 ± 0.008	0.937 ± 0.005	0.897 ± 0.009
Magnitude-based	44.3	41.2	4.9	0.899 ± 0.006	0.947 ± 0.003	0.921 ± 0.006
Divergence-only	44.9	41.9	4.8	0.908 ± 0.005	0.952 ± 0.003	0.935 ± 0.005
Morphology-aware proposed	45.5	42.6	4.7	0.913 ± 0.004	0.955 ± 0.002	0.941 ± 0.004
Attention-enhanced U-Net						
None	0	0	4.5	0.920 ± 0.003	0.958 ± 0.002	0.946 ± 0.003
Random structured	49.5	46.1	3.3	0.882 ± 0.008	0.937 ± 0.005	0.890 ± 0.009
Magnitude-based	49.1	46.0	3.3	0.899 ± 0.005	0.947 ± 0.003	0.911 ± 0.006
Divergence-only	49.4	46.6	3.2	0.907 ± 0.004	0.951 ± 0.003	0.925 ± 0.005
Morphology-aware proposed	50.7	47.3	3.1	0.912 ± 0.004	0.954 ± 0.002	0.936 ± 0.004
Proposed + block pruning	57.8	54.6	2.9	0.906 ± 0.005	0.951 ± 0.003	0.928 ± 0.005

The proposed criterion consistently outperforms random, magnitude-based, and divergence-only pruning across all fourteen models, with the largest gains in BF1, indicating it better preserves contour-relevant components. For the Attention-enhanced U-Net, adding contextual block pruning increases compression further at a modest cost in quality.

Together, these experiments show that combining divergence-based and morphology-aware importance yields a pruning criterion that is more accurate and more consistent across architectures than magnitude- or single-signal alternatives, while remaining efficient for practical use.

VI. Conclusion

This study introduced a morphology-aware divergence-guided structured pruning method for efficient semantic segmentation of streptococci in microscopy images of live, unfixed microorganisms. The proposed approach combines representation-based component importance with foreground- and multi-scale boundary-sensitive activation information, enabling structural redundancy to be reduced while preserving features important for streptococcal regions and their contours. Representation-preserving fine-tuning further mitigates the accuracy loss caused by structural compression.

Experiments on the pixel-level annotated streptococci subset of MMICD demonstrated a favorable accuracy–efficiency trade-off compared with random structured, magnitude-based, and divergence-only pruning. At the selected operating point, the compressed model retained a Dice coefficient of 0.954, an IoU of 0.912, and a Boundary F1 score of 0.936, while reducing the number of parameters by 50.7%, FLOPs by 47.3%, and inference time by 31.1%. The ablation and compression-level analyses further confirmed that morphology-sensitive component selection improves the preservation of segmentation-relevant spatial information under structural compression.

These results demonstrate that the proposed pruning strategy can substantially reduce model complexity with only a minor degradation in segmentation quality, supporting its use in high-throughput and computationally constrained microscopy applications.

References

[1] J. J. Baty, S. N. Stoner, and J. A. Scoffield, "Oral commensal streptococci: Gatekeepers of the oral cavity," *Journal of Bacteriology*, vol. 204, no. 11, p. e0025722, Nov. 2022, pMID: 36286512. PMC: PMC9664950. NLM owner: NLM. Status: MEDLINE. Date completed: 2022-11-18. Last revision: 2024-09-07. Publication types: Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't; Review. Publication place: United States. Conflict of interest: The authors declare no conflict of interest.

[2] M. Han, Y. Wu, X. Guo, L. Jiang, X. Wang, and Z. Gai, "Milk fermentation by monocultures or co-cultures of streptococcus thermophilus strains," *Frontiers in Bioengineering and Biotechnology*, vol. Volume 10 - 2022, 2022.

[3] A. Samarin, A. Savelev, A. Toropov, A. Nazarenko, A. Motyko, E. Kotenko, A. Dozorceva, A. Dzestelova, E. Mikhailova, and V. Malykh, "Streptococci recognition in microscope images using taxonomy-based visual features," *Opt. Mem. Neural Netw.*, vol. 33, no. Suppl 3, p. S424–S434, Dec. 2024. [Online]. Available: https://doi.org/10.3103/S1060992X24700693

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[5] C. Spahn, E. Gómez-de Mariscal, R. F. Laine, P. M. Pereira, L. von Chamier, M. Conduit, M. G. Pinho, G. Jacquemet, S. Holden, M. Heilemann, and R. Henriques, "Deepbacs for multi-task bacterial image analysis using open-source deep learning approaches," *Communications Biology*, vol. 5, no. 1, p. 688, Jul. 2022. [Online]. Available: https://doi.org/10.1038/s42003-022-03634-z

[6] K. J. Cutler, C. Stringer, T. W. Lo, L. Rappez, N. Stroustrup, S. Brook Peterson, P. A. Wiggins, and J. D. Mougous, "Omnipose: A high-precision morphology-independent solution for bacterial cell segmentation," *Nature Methods*, vol. 19, no. 11, pp. 1438–1448, Nov. 2022.

[7] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," *CoRR*, vol. abs/1711.07971, 2017. [Online]. Available: http://arxiv.org/abs/1711.07971

[8] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR*, vol. abs/2102.04306, 2021. [Online]. Available: https://arxiv.org/abs/2102.04306

[9] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *CoRR*, vol. abs/1608.08710, 2016. [Online]. Available: http://arxiv.org/abs/1608.08710

[10] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1398–1406.

[11] A. Samarin, A. Savelev, A. Toropov, A. Dozortseva, E. Kotenko, A. Nazarenko, A. Motyko, G. Narova, E. Mikhailova, and V. Malykh, "Automation of multi-class microscopy image classification based on the microorganisms taxonomic features extraction," *Journal of Imaging*, vol. 11, no. 6, 2025. [Online]. Available: https://www.mdpi.com/2313-433X/11/6/201

[12] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, ser. Lecture Notes in Computer Science, vol. 11045. Switzerland: Springer, Sep. 2018, pp. 3–11.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 833–851.

[14] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: A generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, no. 1, pp. 100–106, Jan. 2021.

[15] A. Samarin, D. Nazarenko, and A. Mamaeva, "Domain adaptation of the segment anything model for monococci segmentation in microscopy images," in *2026 International Russian Smart Industry Conference (SmartIndustryCon)*, 2026, pp. 559–565.

[16] ——, "Grounding dino–based detection of bacilli in microscopy images," in *2026 International Russian Smart Industry Conference (SmartIndustryCon)*, 2026, pp. 566–572.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[18] A. Samarin, E. Kotenko, A. Toropov, A. Nazarenko, A. Savelev, A. Dozortseva, A. Mamaeva, A. Motyko, E. Mikhailova, and V. Malykh, "Refined non-local blocks for precise segmentation of diplococci in microscopy imagery," in *Proceedings of the 2026 9th International Conference on Information and Computer Technologies*, ser. ICICT 2026. New York, NY, USA: Association for Computing Machinery, 2026, p. 155–161.

[19] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient transfer learning," *CoRR*, vol. abs/1611.06440, 2016. [Online]. Available: http://arxiv.org/abs/1611.06440

[20] C. I. López-González, E. Gascó, F. Barrientos-Espillco, E. Besada-Portas, and G. Pajares, "Filter pruning for convolutional neural networks in semantic image segmentation," *Neural Netw.*, vol. 169, no. C, p. 713–732, Jan. 2024.

[21] J. M. Valverde, A. Shatillo, and J. Tohka, "Sauron u-net: Simple automated redundancy elimination in medical image segmentation via filter pruning," *Neurocomputing*, vol. 594, p. 127817, Aug. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2024.127817

[22] D. Wu, Z. Guo, L. Yu, N. Sang, and C. Gao, "Structural pruning via spatial-aware information redundancy for semantic segmentation," in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025.

[23] A. Samarin, A. Nazarenko, E. Kotenko, A. Savelev, A. Toropov, A. Motyko, and V. Malykh, "Idap++: Advancing divergence-aware pruning with joint filter and layer optimization," *Proc. ACM Manag. Data*, vol. 4, no. 1, Apr. 2026.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[25] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[26] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 6881–6890.

[27] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 12 077–12 090. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf

[28] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 1290–1299.

[29] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," 2021. [Online]. Available: https://arxiv.org/abs/2105.05537

[30] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," 2021. [Online]. Available: https://arxiv.org/abs/2106.04803

[31] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.

[32] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," in *International Conference on Learning Representations*, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, Eds., vol. 2025, 2025, pp. 28 085–28 128. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2025/file/45c1f6a8cbf2da59ebf2c802b4f742cd-Paper-Conference.pdf

[33] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, Jan. 2024. [Online]. Available: http://dx.doi.org/10.1038/s41467-024-44824-z