applied sciences

IMPACT FACTOR 2.9

CITESCORE 6.1

Article

Development and Evaluation of YOLO26-Refined: A P2-FPN and CBAM-Enhanced Architecture for Detection of Gram-Stained Bacterial Cells in Microscopic Images

Dmitriy Berillo, Kainizhamal Iklassova, Rustem Tashibayev, Gulnar Kim, Vladislav Semenyuk, Ildar Kurmashev and Danila Zhirnov

MDPI

https://doi.org/10.3390/app16168289

applied sciences

MDPI

Article

Development and Evaluation of YOLO26-Refined: A P2-FPN and CBAM-Enhanced Architecture for Detection of Gram-Stained Bacterial Cells in Microscopic Images

Dmitriy Berillo, Kainizhamal Iklassova, Rustem Tashibayev *, Gulnar Kim *, Vladislav Semenyuk, Ildar Kurmashev and Danila Zhirnov

Non-Profit Limited Company "Manash Kozybayev North Kazakhstan University", Petropavlovsk City 150000, Kazakhstan; daberillo@ku.edu.kz (D.B.); vvsemenyuk@ku.edu.kz (V.S.)
* Correspondence: rustashibaev@gmail.com (R.T.); gakim@ku.edu.kz (G.K.); Tel.: +7-775-213-40-64 (R.T.); +7-707-117-77-35 (G.K.)

Abstract

Gram staining remains a primary diagnostic method in clinical microbiology, but manual interpretation of stained smears is time-consuming, subject to inter-operator variability, and impractical for continuous 24/7 monitoring. Automated detection offers faster turnaround, standardized classification independent of operator experience, and scalability for routine water-quality surveillance. This study evaluates seven YOLO-based architectures for four-class detection of Gram-stained bacterial cells (Gram-positive cocci, Gram-negative cocci, Gram-positive rods, Gram-negative rods) in microscopic images at 1000× magnification. Four baseline nano-scale models (YOLOv10n, YOLO11n, YOLO12n, YOLO26n) were trained and compared with three modified architectures based on YOLO26, including the proposed YOLO26-Refined, which extends the feature pyramid to the P2/4 level, applies a CBAM attention module after the P3 stage, disables shortcut connections in neck blocks, and uses reg_max = 16 for Distribution Focal Loss. All models were trained on a dataset of 5994 annotated images (69/21/10 split) using an NVIDIA RTX 4080 GPU. YOLO26-Refined achieved the best overall performance, with mAP50 = 0.934, mAP50-95 = 0.616, Precision = 0.846, and Recall = 0.888, outperforming the YOLO26n baseline (mAP50 = 0.923) by 0.011. Per-class results show the highest accuracy for Gram-negative rods (mAP50 = 0.965) and the lowest for Gram-positive rods (mAP50 = 0.890). With 2.63 million parameters, 9.9 GFLOPs, and an inference latency of approximately 1.8 ms per image on an RTX 4080 GPU (batch = 1, FP16), the model remains suitable for edge deployment, supporting its application in automated water-quality microbiological monitoring under the IRN BR28712227 research grant.

Keywords: Gram stain; bacteria detection; YOLO; deep learning; object detection; CBAM; feature pyramid network; microscopy image analysis; clinical microbiology; edge deployment

Check for updates

Academic Editor: Atsushi Mase

Received: 17 July 2026
Revised: 13 August 2026
Accepted: 17 August 2026
Published: 20 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Bacterial infectious diseases remain one of the leading causes of morbidity and mortality worldwide, requiring rapid and accurate laboratory diagnostics for the timely administration of antibacterial therapy [1]. Gram staining, developed by Hans Christian Gram in 1884, remains one of the most widely used primary diagnostic methods in clinical microbiology: within a few minutes, it makes it possible to divide bacteria into Gram-positive

and Gram-negative groups, to differentiate cocci from rods, and to obtain preliminary information about the pathogen needed for the choice of empirical antibiotic therapy [2,3].

Traditional interpretation of Gram-stained smears is performed manually by a qualified microbiologist. This process is labor-intensive, subjective, and subject to inter-operator variability caused by differences in staining protocols, slide preparation, and individual interpretation [4]. Automation of Gram-slide interpretation can substantially reduce turnaround time and improve the reproducibility of diagnostics [1].

Alternative methods of bacterial identification, such as mass spectrometry (MALDI-TOF), PCR, and next-generation sequencing, provide high accuracy down to the species level; however, they require specialized equipment and considerable time and financial costs, which limits their use in resource-constrained institutions [5]. Gram staining combined with automated image analysis retains a low cost and procedural simplicity while adding the speed and objectivity of computer-based analysis.

Deep learning, and in particular, convolutional neural networks (CNNs), demonstrates high efficiency in biomedical image analysis tasks. A number of studies have applied the classical ResNet, VGG, and DenseNet architectures to the classification of bacterial morphology; however, most such approaches are oriented toward the classification of individual cells rather than object detection, with spatial localization across the full microscope field of view [6,7]. The detection task is fundamentally more complex, since the model must simultaneously localize and classify a large number of small objects under conditions of high cell density and variable slide background. A particular challenge arises from the morphological overlap between classes: Gram-positive and Gram-negative cocci share a similar spherical shape, while rod-shaped cells of both staining groups can appear nearly identical when partially overlapping or when stain uptake varies across the smear. These subtle inter-class similarities make four-class detection substantially harder than binary Gram-positive/Gram-negative classification.

The YOLO family of algorithms has established itself as the de facto standard for real-time object detection tasks owing to its optimal balance between accuracy and inference speed [8]. Beginning with YOLOv1 [9], each successive generation—YOLOv3 [10], YOLOv5 [11], YOLOv8 [12], YOLOv9 [13]—introduced architectural improvements to feature-extraction mechanisms, the Feature Pyramid Network (FPN) [14], and loss functions. Modern architectures such as YOLOv10 [15], YOLO11 [16], YOLO12 [17], and YOLO26 [18] additionally introduce attention mechanisms, area attention, and end-to-end detection without NMS, expanding the possibilities of application in biomedical diagnostics. The deployment of deep-learning models on resource-constrained edge devices, such as single-board computers connected to laboratory microscopes, imposes strict limits on model size and computational cost, making nano-class architectures a practical requirement for real-time clinical microscopy.

The importance of automating microbiological analysis is confirmed in practice. Under the program-targeted funding of the Ministry of Science and Higher Education of the Republic of Kazakhstan (project IRN BR28712227, 2025–2027), the NPJSC M. Kozybayev North Kazakhstan University (Petropavlovsk) is implementing a program for the development of high-tech solutions for the monitoring and analysis of water resources of the North Kazakhstan Region, including the creation of a laboratory of multimodal microscopy and biosensor technologies. Automated detection and identification of bacteria by the Gram stain method is a key tool of sanitary-epidemiological water-quality control in this project.

Despite the advances achieved, the task of highly accurate detection of small bacterial cells on Gram-stained slides under conditions of limited computational resources remains open. Most existing approaches do not consider four-class detection with simultaneous separation by cell shape (cocci/rods) and staining result (Gram-positive/Gram-negative),

nor do they take into account the requirements for deployment on edge devices. The use of lightweight nano-class architectures with CBAM attention mechanisms [19] and an extended P2–P5 feature pyramid represents a promising direction for solving this task.

The main contributions of this work are as follows: (1) a systematic benchmarking of four baseline YOLO architectures (YOLOv10n, YOLO11n, YOLO12n, YOLO26n) for four-class Gram-stain bacterial detection on a unified dataset; (2) the design and evaluation of three progressively modified YOLO26 variants (YOLO26-Bacteria, YOLO26-Hybrid, YOLO26-Refined) with ablation analysis of each architectural component; (3) a quantitative bounding-box size analysis justifying the extension of the feature pyramid to the P2/4 level; (4) the assembly of a multi-source dataset of 5994 annotated microscopic images from three independent laboratories. The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset, Gram staining protocol, baseline and proposed architectures, and training configuration. Section 4 presents the experimental results. Section 5 provides a discussion of the findings, including ablation analysis, per-class evaluation, comparison with published approaches, and limitations. Section 6 summarizes the conclusions and outlines future research directions.

2. Related Work

Automated recognition of bacterial morphology in microscopic images is an active area of research in medical diagnostics. The first study to apply deep learning to the automated interpretation of Gram-stained blood smears was the work of Smith et al. [7]. In that study, a CNN was trained on 100,213 crops from 25,488 images of positive blood cultures. Sensitivity for Gram-positive cocci in clusters was 93.2%, and for Gram-negative rods, it was 96.3%. Kristensen et al. [20] developed a system for the classification of Gram-stained microscopic images based on image processing and automated classification models, demonstrating the feasibility of reliable automated interpretation with a limited volume of training data. Kim et al. [21] investigated the use of rapid CNNs on mobile devices for Gram-stain classification in real-time inference mode. Hata et al. [22] proposed an approach to Gram-stain quality control using a CNN as a standardized QC method.

In parallel, the direction of transformer architectures for Gram-stain analysis has been developed. Walter et al. [23] proposed GramViT—a vision-transformer-based approach for automating the analysis of whole-slide images (WSI) that eliminates the need for manual annotation. Walter et al. [1] developed a clinical-grade CNN tool for the characterization of Gram-stained blood-culture slides, evaluated by qualified microbiologists. Signoroni et al. [24] presented a hierarchical AI for the global interpretation of culture plates in the era of digital microbiology.

In bacterial detection tasks with localization, architectures of the YOLO family have found wide application. Chin et al. [25] systematically analyzed machine- and deep-learning-based methods for bacterial detection and classification as applied to clinical microscopy; it was shown that YOLO-based methods provide the best balance of accuracy and inference speed among single-stage detectors. Sun et al. [26] applied YOLO for the long-term tracking and detection of *E. coli* in microchannels, achieving mAP = 94.7% at 1000× magnification. Sroka-Oleksiak et al. [27] proposed an AI-based approach for the identification of bacterial and fungal pathogens in blood smears. Xu et al. [28] developed TE-YOLOF—a compact detector for blood cells that achieved a mAP = 90.9%, with a substantially smaller number of parameters compared with those required for the baseline YOLOF.

An important auxiliary direction is the segmentation of individual bacterial cells. Panigrahi et al. [29] developed MISIC—a deep-learning-based segmentation method for various bacterial morphologies with high throughput. Jelli et al. [30] created a

crowdsourcing annotation tool for single-cell segmentation. Mantes et al. [31] proposed MuViT—multi-resolution vision transformers for learning accurate low-resolution representations in microscopy. Darnige et al. [32] developed a deep-learning-based method for single-cell tracking in microscopy. Li et al. [33] used deep learning to detect hyperactivated platelets in whole blood. Iida et al. [7] developed a system for detecting bacterial Gram types using deep learning.

The development of YOLO architectures follows a path of consistent improvement of feature-extraction mechanisms and detection objects. YOLOv1 [9] laid down the concept of a single pass for simultaneous localization and classification. YOLOv2 [34] introduced anchor boxes and batch normalization, and YOLOv3 [10] introduced three-scale detection with an FPN-like structure. YOLOv4 [35] systematized training techniques (bag-of-freebies) and achieved a 43.5% AP on MS COCO at ~65 FPS on a Tesla V100. YOLOv5 [11] standardized a modular PyTorch implementation. YOLOv6 [36] and YOLOv7 [37] extended the range of speed and accuracy for industrial applications. YOLOv8 [12] introduced a decoupled detection head and anchor-free prediction. YOLOv9 [13] introduced GELAN and progressive knowledge distillation. YOLOv10 [15] eliminated the need for NMS through dual target assignment. YOLO11 [16,38] improved GPU efficiency by introducing C3k2 and C2PSA blocks. YOLO12 [17] introduced the area-attention mechanism and R-ELAN, achieving mAP50-95 = 40.6% with a latency of 1.64 ms on a T4 GPU. YOLO26 [18,39] implemented the most radical simplifications: the removal of DFL [40] and NMS and the introduction of MuSGD, ProgLoss, and Small-Target-Aware Label Assignment (STAL).

The Feature Pyramid Network (FPN) [14] is a key component of modern detectors, providing multi-level feature representation with minimal computational cost. A standard FPN detects objects at the P3/8, P4/16, and P5/32 levels. Extension to the P2/4 level (160 × 160 pixels) is critically important for the detection of small objects—in particular, small bacterial cells on Gram slides, which occupy only a few pixels in a 640 × 640 image. The CBAM attention mechanism [19], proposed by Woo et al., sequentially applies channel and spatial attention to feature maps, which demonstrates a stable improvement in the quality of detection and classification across various CNN architectures, including biomedical tasks.

Table 1 provides a comparative overview of published works on the application of YOLO architectures for bacterial detection. The analysis shows that, despite significant progress in general object-detection tasks, the task of multi-class Gram-stain bacterial detection, with simultaneous separation by cell shape and staining result under conditions of limited computational resources, remains poorly studied. The present work aims to fill this gap through the development and systematic evaluation of a family of architectures based on YOLO26 with an extended P2-FPN and a CBAM attention mechanism.

Most published works address the classification task without object localization or are limited to a binary GP/GN separation. Four-class detection with separation by cell shape and staining result under conditions suitable for edge devices has essentially not been considered in the literature.

(1) Preparation of a Gram-stain dataset with four object classes: Gram-positive cocci, Gram-negative cocci, Gram-positive rods, and Gram-negative rods. The dataset is formed on the basis of public image sets that were split and annotated on open cloud platforms, as well as on the basis of the authors' own samples obtained during experiments. It is envisaged to use the Roboflow service for additional processing of the dataset, including filtering of incorrect images, correction of annotation, augmentation, and changes to the data format.

(2) Training of the current versions of YOLO detectors available in the Ultralytics library: YOLOv10n, YOLO11n, YOLO12n, and YOLO26n. In this work, the "nano–n" model

was chosen, which is dictated by the requirements of deploying the product on low-performance "edge" devices, such as microcomputers and industrial PCs with access to the microscope video stream (with software of the ToupCam type). The use of heavier architectures (s/m/l/x) may prove unacceptable for real-time operation.

(3) Development of the authors' own architectures (modifications of YOLO26n) of visual detectors aimed at improving the performance metrics, i.e., mean average precision mAP50 and mAP50-95, Precision, Recall, and the F1 metric.

(4) Training, validation, and testing of the developed modifications as improved versions of the YOLO26n architecture.

(5) Performing a comparative analysis of the obtained metrics and ablation studies to assess the contribution of each architectural component. When signs of overfitting or underfitting are detected, the hyperparameters and the data-augmentation strategy are adjusted to achieve the optimal condition of the model.

(6) Formulation of a conclusion on the effectiveness of the developed model for the task of bacterial detection by Gram staining, determination of the limitations of the proposed approach, and identification of directions for further development, in particular the expansion of the dataset and adaptation to new or heavier versions of YOLO.

Table 1. Comparative overview of deep-learning methods for bacterial detection and classification. Metrics are given for the best configuration of each model according to the data of the original publications.

Reference	Method/Model	Task	Object/Dataset	Accuracy/Sensitivity	Recall/Specificity	mAP/AUC
[7]	Deep CNN (VGG-style)	Gram-stain classification (WSI)	Blood cultures; 25,488 images; 189 slides	Accuracy 94.9% (crops); 92.5% (slides)	Sensitivity 93.2–98.4%; Specificity 75.0–98.1%	AUC > 0.98 for each class
[20]	Random Forest + CPN	Gram-stain classification	660 images; 33 species of bacteria/fungi	Accuracy of RF: 99%; CPN: 80%	GP/GN: RF 99%; CPN 95%	n/a
[21]	MobileNet/Inception3/ ResNet50	Gram-stain classification (mobile devices)	MHU dataset; clinical slides	~90% (baseline)	n/a	n/a
[25]	Deep CNN (Cognex ViDi Suite)	Gram-stain quality control (QC)	105 training + 92 test images	Accuracy 100%; Recall 100%	n/a	F1 = 1.00
[28]	YOLOv4 (best of 3 models)	Detection + classification of *E. coli* (three growth stages)	*E. coli*; microscopic images	Precision = 86%	Recall = 97%	mAP = 98%; F1 = 91%
[31]	YOLOv5 (with positioning microspheres)	Detection + tracking of *E. coli* in microchannels	*E. coli*; microfluidic chip; oil 1000×	Accuracy of bacteria: 0.734; microspheres: 0.910	Recall of microspheres: 0.927	n/a
[32]	Cellpose 3 + Attention MIL	Detection + classification of 14 bacteria + 3 fungi	Blood smears of septic patients; 16,637 images	Accuracy of bacteria: 77.15%; fungi: 71.39%	Best species (*C. acnes*, *E. faecium*): up to 96.2%	AUC 0.97 (bacteria); AUC 0.88 (fungi)
[14]	TE-YOLOF (EfficientNet backbone)	Blood cell detection	BCCD dataset (erythrocytes, leukocytes, platelets)	n/a	n/a	mAP = 91.9% on BCCD

GP = Gram-positive; GN = Gram-negative; WSI = whole-slide image; QC = quality control; CNN = convolutional neural network; VGG = visual geometry group; RF = random forest; MIL = multiple instance learning; BCCD = blood cell count and detection; MHU = mie university hospital, CPN = color pixel network; YOLO = you only look once; mAP = mean average precision; AUC = area under the ROC curve; n/a = not specified in the original work. The work [14] addresses the task of blood-cell detection, not Gram staining.

3. Materials and Methods

3.1. Dataset

The experimental dataset was obtained from multiple sources to ensure morphological diversity and cross-laboratory representativeness. The basis was composed of images from the open Zenodo repository [41], which contains annotated bright-field micrographs of clinical bacterial smears. To expand the dataset across all four classes, additional microscopic images were obtained from three independent laboratories: (1) the microbiological laboratory of M. Kozybayev North Kazakhstan University (Petropavlovsk, Kazakhstan); (2) the Sanitary-Epidemiological Station (SES) of Petropavlovsk; and (3) Al-Farabi Kazakh National University (Almaty, Kazakhstan). The inclusion of images from laboratories in two different cities with different equipment and staining protocols enhances the external validity of the trained models.

The dataset covers four object classes: Gram-positive cocci, Gram-negative cocci, Gram-positive rods, and Gram-negative rods. Imaging was performed at 1000× magnification in immersion oil in bright-field mode; the images were resized to 640 × 640 pixels with automatic orientation correction. The dataset was split into training (69%, 4129 images), validation (21%, 1271 images), and test (10%, 594 images) subsets, for a total of 5994 images. The split was performed in a stratified manner. Annotation quality was ensured through a multi-stage verification process: initial bounding-box annotations were generated using the Roboflow labeling interface (Roboflow, Inc., Des Moines, IA, USA; web-based platform), followed by manual review by two independent annotators with microbiology expertise. Disagreements were resolved by consensus, and images with ambiguous or low-quality staining were excluded from the dataset.

To quantify the object-size distribution in the dataset, the maximum bounding-box side was computed for all 12,278 annotated objects across the four classes (Figure 1). The median maximum side is 56 pixels (mean 58.4 px) at 640 × 640 resolution, with the 5th–95th percentile range spanning 40–85 px. According to the standard MS COCO size taxonomy, 97.1% of the objects fall in the medium category (32–96 px), 0.7% are small (<32 px), and 2.2% are large (>96 px). Coccal classes exhibit smaller typical sizes (GN cocci: median 49 px; GP cocci: 47 px) compared with rod-shaped classes (GN rods: 63 px; GP rods: 60 px). Although the majority of objects are formally classified as medium, they occupy only 6–7 cells on the P3/8 feature map (80 × 80 resolution, stride 8). Extension to P2/4 (160 × 160, stride 4) quadruples the spatial resolution, providing 12–14 feature-map cells per median-sized object, which substantially improves bounding-box localization accuracy as reflected in the mAP50-95 metric.

Augmentation was not applied during data preparation; all transformations were performed during training (Section 3.5).

The conditions of the laboratory experiments at Kozybayev University are documented in Figure 2.

3.2. Gram-Staining Method

The Gram-staining procedure was carried out as follows. From a pure culture grown on nutrient agar for 18–24 h, a thin smear was prepared on a glass slide, fixed by passing it over an open flame, and allowed to cool. The smear was flooded with crystal violet solution for 60 s, rinsed with distilled water, and treated with Lugol's solution for 60 s. Decolorization was performed with 95% ethanol for 15–30 s, after which the slide was immediately rinsed with water [4]. Counterstaining with safranin was performed for 60 s. After the final rinse and drying, the slides were examined at 1000× magnification with immersion oil.

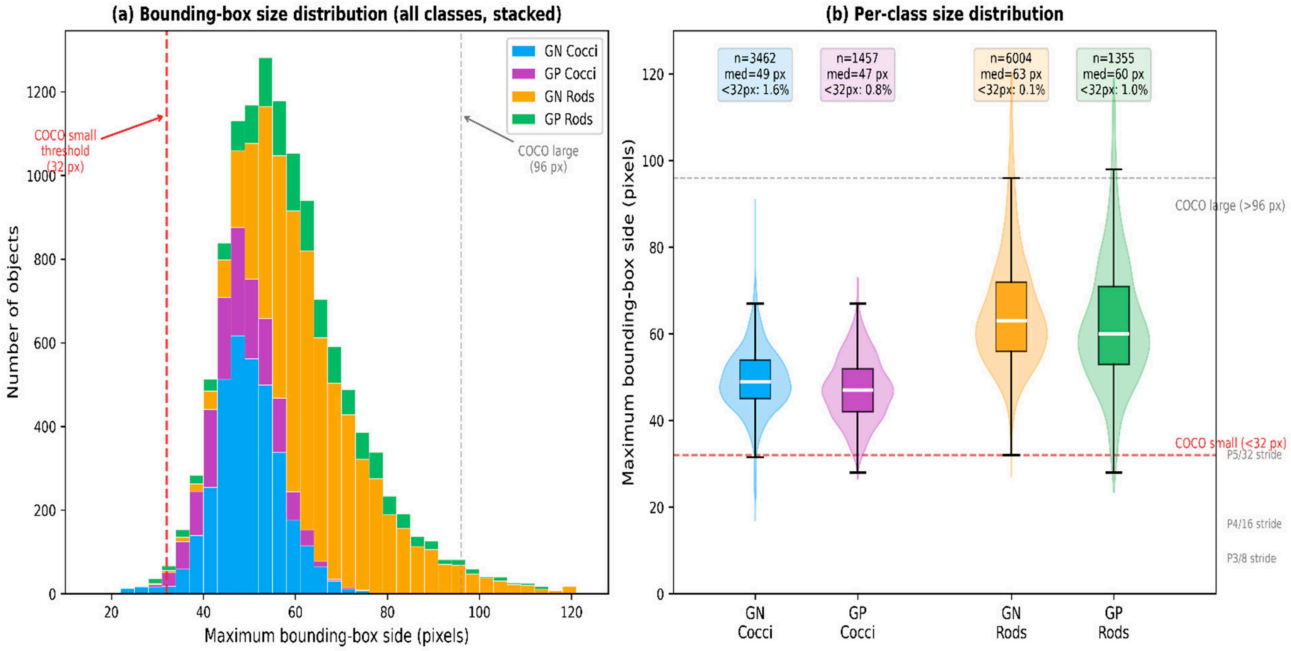


Figure 1. Distribution of bounding-box sizes in the dataset (12,278 objects, 640 × 640 pixels): (**a**) stacked histogram of the maximum bounding-box side across all four classes with COCO size category thresholds; (**b**) per-class violin and box plots showing the median, interquartile range, and overall distribution shape.

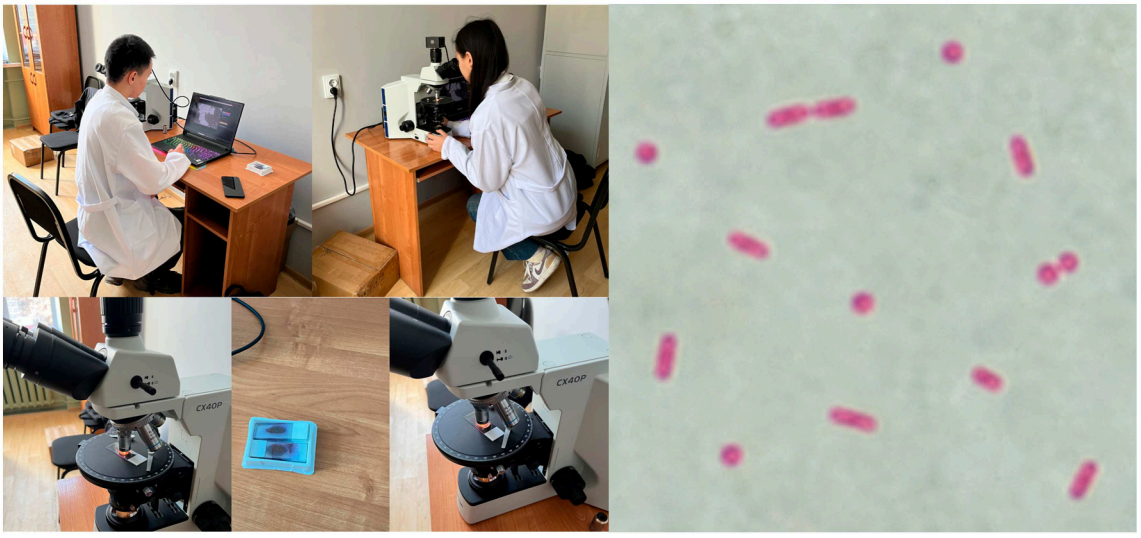


Figure 2. Laboratory experiments in the microbiological laboratory of Kozybayev University (Petropavlovsk, Kazakhstan): preparation of Gram-stained smears for the formation of the dataset.

Gram-positive microorganisms retain crystal violet and appear violet—the thick peptidoglycan layer prevents the complex from being washed out during decolorization. Gram-negative microorganisms lose the primary dye, since the thin peptidoglycan layer and the lipopolysaccharide-containing outer membrane are disrupted during decolorization; these cells take up safranin and appear pink-red [2,3]. The four morphological classes considered in this work differ both in staining result and in cell shape, which makes the task of automated detection a nontrivial multi-class object-localization problem [1].

3.3. Baseline YOLO Architectures

To evaluate performance, four current architectures from the Ultralytics library were selected as baseline models, representing the current level of lightweight "n"-level detectors.

3.3.1. YOLOv10n

YOLOv10 [15] presents a two-branch detection architecture that eliminates the need for Non-Maximum Suppression (NMS) through the introduction of consistent dual assignments. The backbone of YOLOv10n (see Figure 3) includes C2f blocks with residual connections, compressed convolutional SCDown modules for downsampling, and a PSA (Partial Self-Attention) module at the top of the feature pyramid. The head uses C2fCIB blocks—a compact inverted narrow-bottleneck structure. The v10Detect detector operates with three FPN levels: P3/8, P4/16, and P5/32. The number of parameters in the "n" version is 2.26 million at 6.5 GFLOPs.

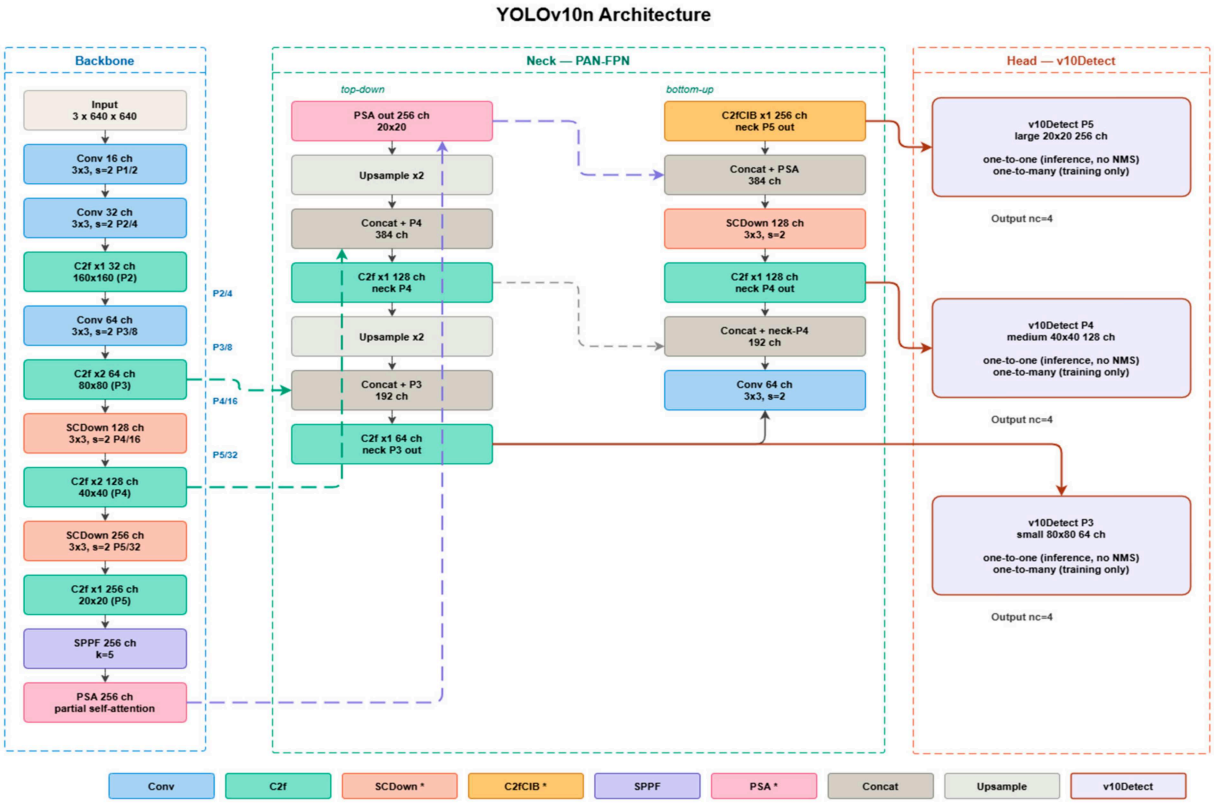


Figure 3. YOLOv10n architecture. The asterisk (*) denotes blocks with configuration parameters specified within the corresponding blocks. Colors indicate different module types as shown in the legend.

3.3.2. YOLO11n

YOLO11 [16] introduces the C3k2 block as an improved variant of C2f, with flexible control of the kernel width and two configuration options: with shortcut connections (True) and without them (False). The backbone of YOLO11n (see Figure 4) sequentially applies C3k2 blocks with an increasing number of channels (256 → 512 → 1024), ending with a standard SPPF(1024, 5) and a C2PSA module that implements a positional self-attention mechanism for global context. The three-level FPN head (P3, P4, P5) forms features through C3k2 blocks, providing 2.58 million parameters and 6.3 GFLOPs while detecting three scales.

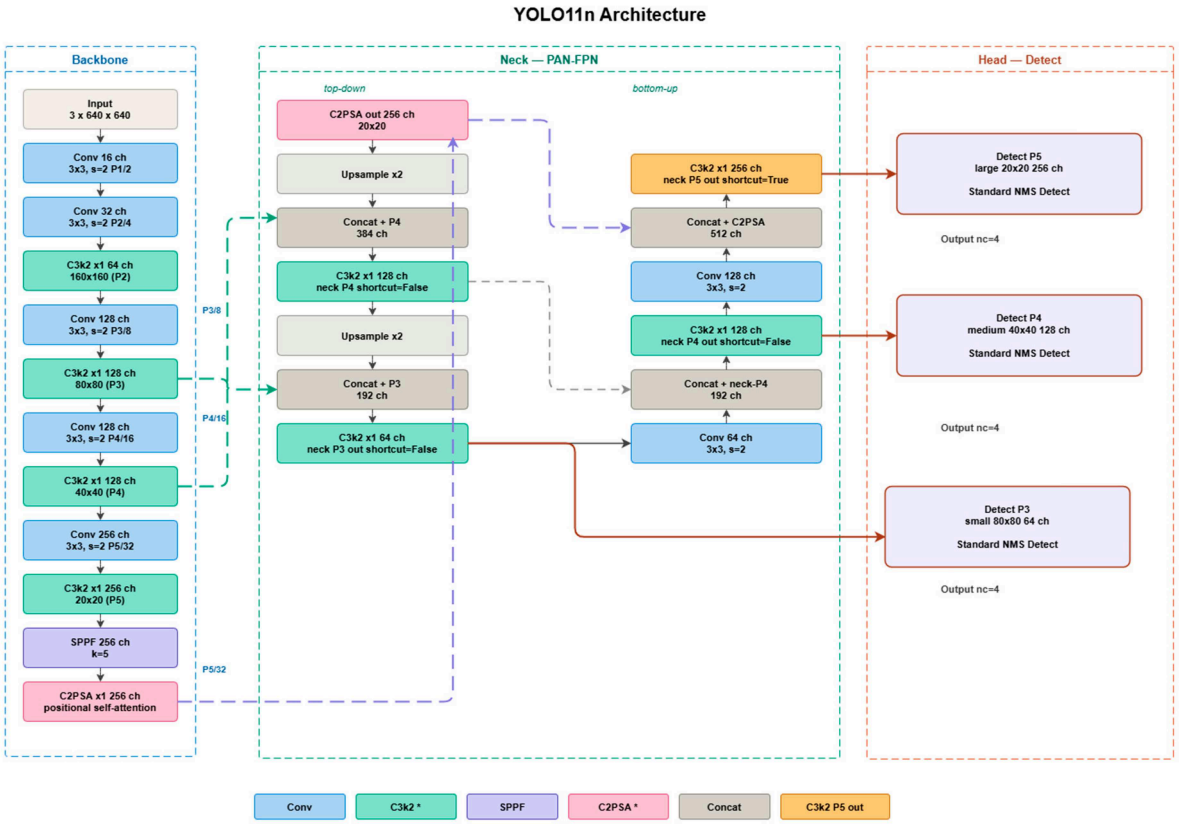


Figure 4. YOLO11n architecture. The asterisk (*) denotes blocks with configuration parameters specified within the corresponding blocks. Colors indicate different module types as shown in the legend.

3.3.3. YOLO12n

YOLO12 [17] introduces the A2C2f (Area Attention C2f) module, which is an attention mechanism with a bounded region that reduces computational complexity compared with that of full self-attention while preserving a global receptive field. In the backbone of YOLO12n (see Figure 5), A2C2f blocks replace part of the C3k2 blocks at the P4 and P5 levels, with attention-head parameters of 4 for P4 and 1 for P5. The head also uses A2C2f for the P3 and P4 levels, retaining C3k2 for P5. The architecture contains 2.55 million parameters at 6.3 GFLOPs. The absence of SPPF and C2PSA in favor of A2C2f is a fundamental architectural difference from YOLO11.

3.3.4. YOLO26n

YOLO26 (see Figure 6) is the original authors' baseline architecture from which all models developed in this work are derived. The backbone of YOLO26n is identical to that of YOLO11n (C3k2, SPPF, and C2PSA blocks); however, it uses an extended SPPF with additional arguments (kernel_size = 5, padding = 3, max_pool = True), which increases the size of the receptive field at the top of the pyramid. The three-level FPN head (P3, P4, P5) is analogous to that of YOLO11n except for the P5 C3k2 block, for which non-standard expansion arguments (e = 0.5, True) are applied. The model uses the end2end = True mode with reg_max = 1 for direct prediction, without NMS. The baseline configuration contains 2.37 million parameters and 5.2 GFLOPs.

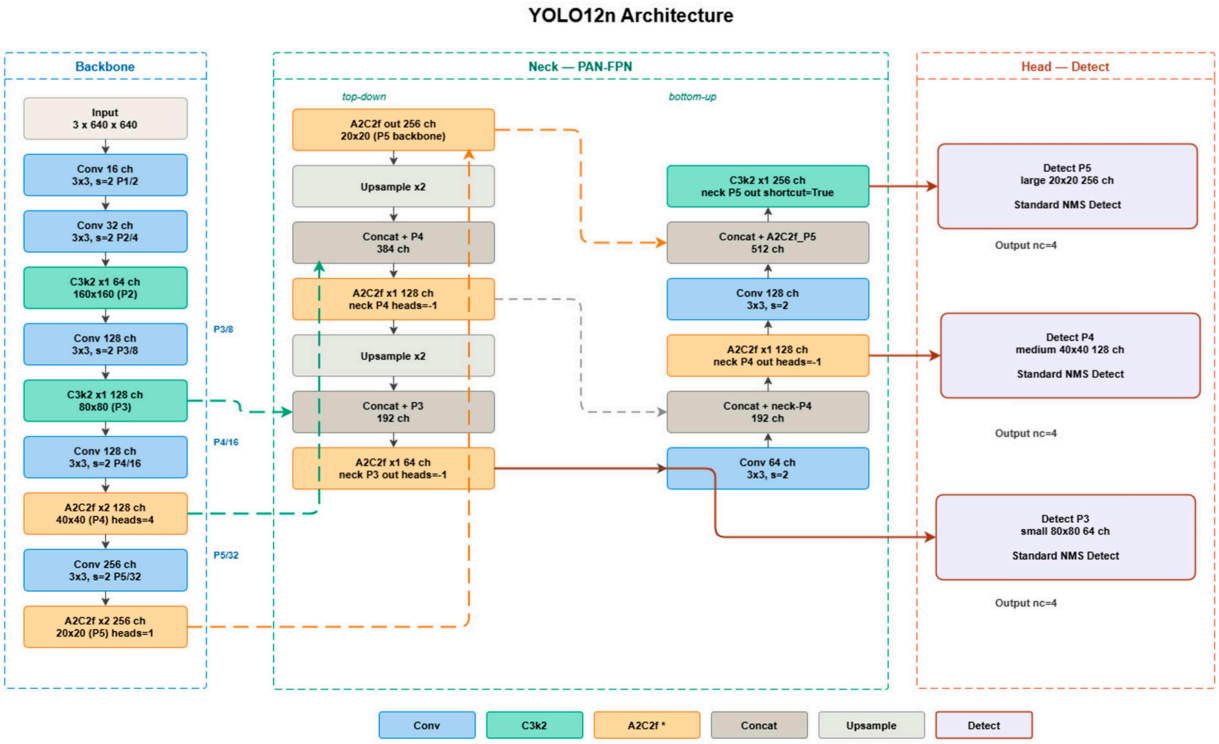


Figure 5. YOLO12n architecture. The asterisk (*) denotes an A2C2f block with configuration parameters specified within the corresponding block. Colors indicate different module types: blue, Conv; green, C3k2; yellow, A2C2f; gray, Concat; beige, Upsample; outlined boxes, Detect.

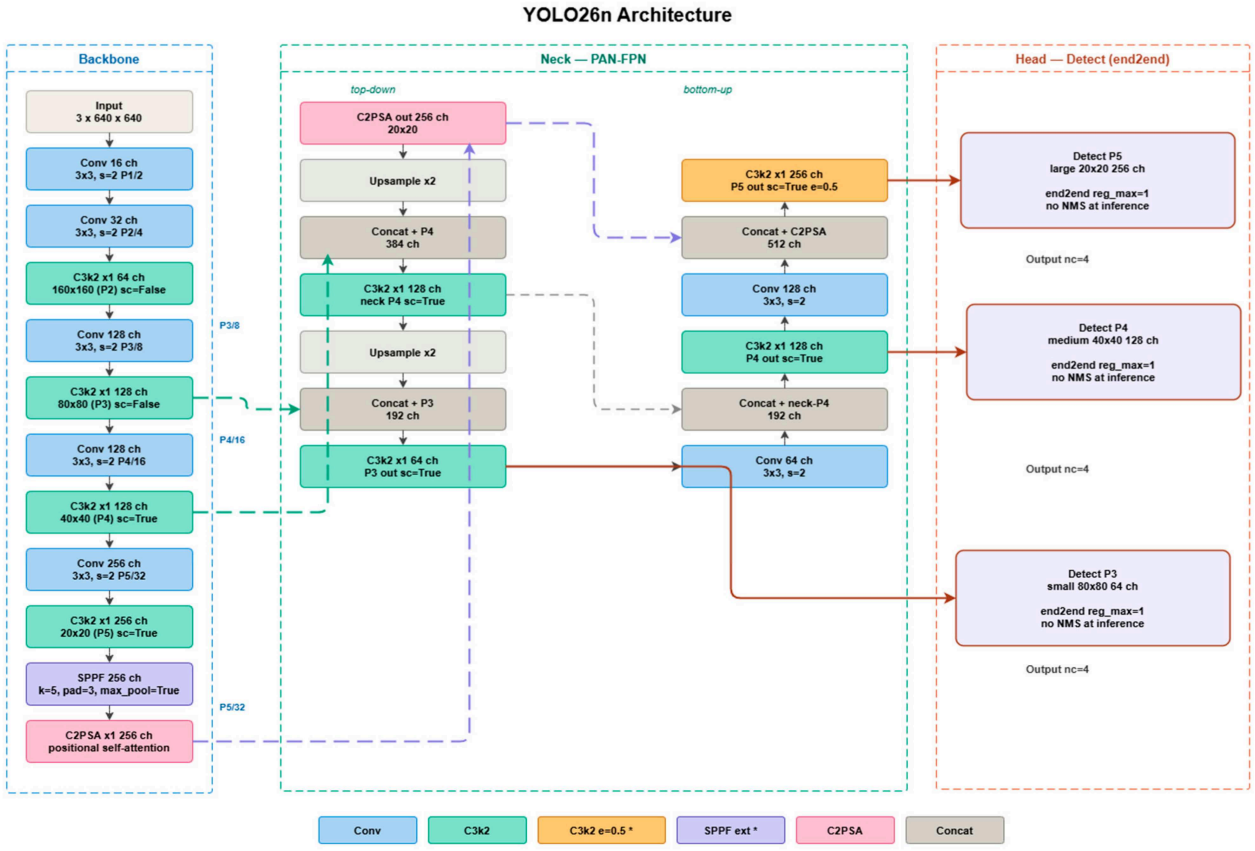


Figure 6. YOLO26n architecture (baseline). The asterisk (*) denotes blocks with configuration parameters specified within the corresponding blocks. Colors indicate different module types as shown in the legend.

3.4. Developed Architectures

Based on an analysis of the specific features of the bacterial-cell detection task, three authors' modifications of the baseline YOLO26n architecture were proposed. The common motivation is improving localization precision for the detected objects. As shown in the bounding-box analysis (Figure 1), bacterial cells have a median size of 56 pixels (47–63 px across classes), occupying only 6–7 cells on the standard P3/8 feature map. Extension to the P2/4 level (160 × 160) doubles the spatial sampling of each object, which is expected to improve bounding-box accuracy.

3.4.1. YOLO26-Bacteria (P2-FPN)

The YOLO26-Bacteria model (see Figure 7) extends the standard three-level FPN of YOLO26n to a four-level P2–P5 pyramid (P2/4, P3/8, P4/16, P5/32). The backbone remains unchanged. A top-down branch is added to the head: after the P3 level, an additional Upsample is performed with concatenation with the P2/4-level features from the backbone (layer 1). A twofold C3k2(128, False) processes the P2 features to obtain a small-object detector. The bottom-up branch builds P3 refined, P4, and P5 with C3k2 blocks, with P5 using a non-standard C3k2(1024, True, e = 0.5) block. The Detect detector receives features of four levels: [P2, P3, P4, P5]. The architecture contains 147 layers and 2.40 million parameters at 6.4 GFLOPs.

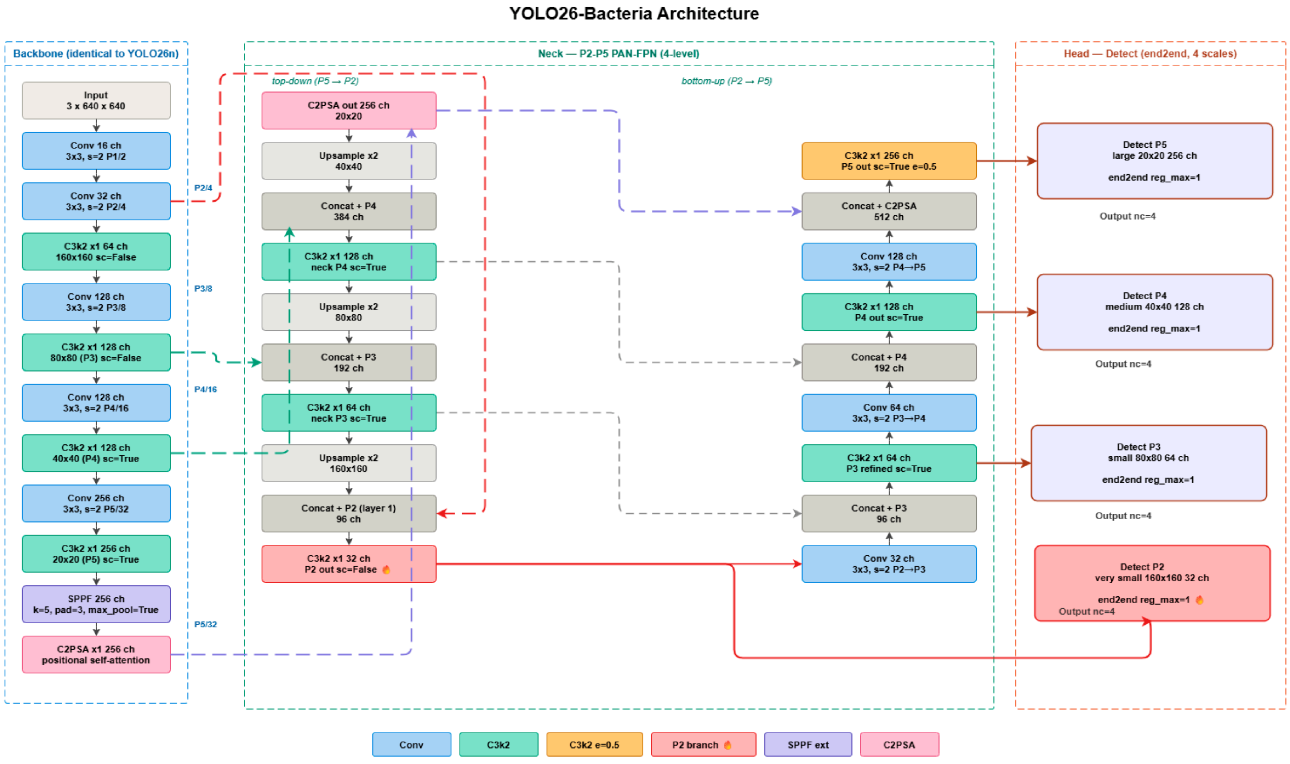


Figure 7. YOLO26-Bacteria architecture with a four-level P2-FPN [9,12,16,17,20,26,31].

3.4.2. YOLO26-Hybrid

The YOLO26-Hybrid model (see Figure 8) is a hybrid architecture that combines the YOLO26n backbone with elements of the YOLO11n head. The backbone is fully identical to that of YOLO26n (including the extended SPPF). The head retains the four-level P2-FPN structure of YOLO26-Bacteria, but differs critically in the following features: all C3k2 blocks in the top-down branch (levels P4 and P3) use shortcut = True instead of False, which strengthens the gradient flow; the P2 level is processed by a single C3k2(128, False) without repetition (repeats = 1) to reduce capacity; the P3 refined level uses a single C3k2(256, True);

and the P5 level applies C3k2(1024, True, e = 0.5) with an additional expansion flag. The Detect detector receives features [P3 refined, P4, P5], excluding P2 from the final detection while retaining P2 refinement in the feature flow. This design choice serves as a deliberate ablation step: the P2 branch acts as an indirect feature enrichment mechanism, where high-resolution P2 features are fused with P3 through the top-down concatenation path, thereby enriching the P3 representation with fine-grained spatial detail without adding a separate detection head at the P2 level. By comparing YOLO26-Bacteria (P2 in Detect) with YOLO26-Hybrid (P2 excluded from Detect), it is possible to isolate the contribution of indirect P2 enrichment vs. direct P2 detection, which informs the final architecture design of YOLO26-Refined. The architecture contains 139 layers and 2.41 million parameters at 6.0 GFLOPs.

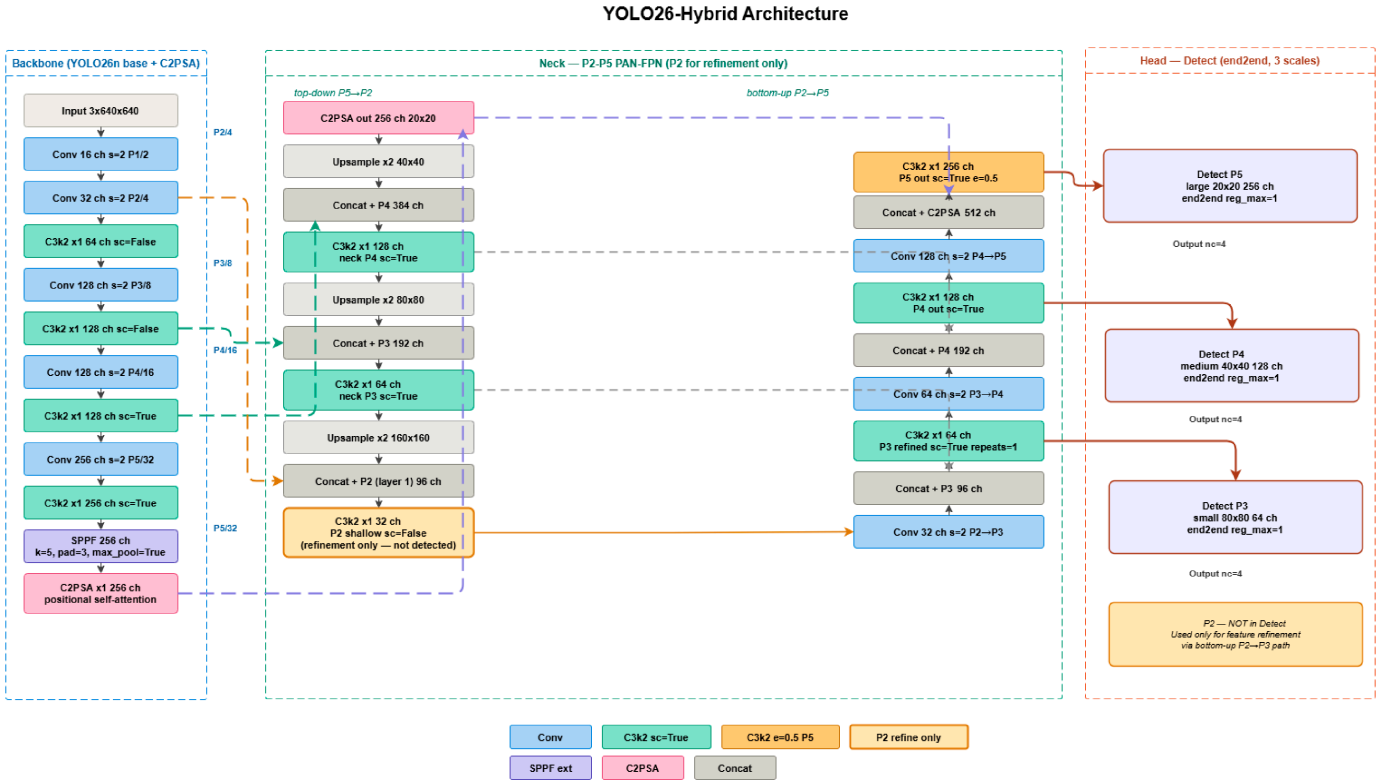


Figure 8. YOLO26-Hybrid architecture [9,12,16,17,20,26,31].

3.4.3. YOLO26-Refined

YOLO26-Refined (see Figure 9) is the final and most effective architecture developed in this study. It consolidates the most effective solutions identified through the sequential ablation analysis of the preceding models. The backbone is identical to that of YOLO26n with one change: the SPPF module is returned to the standard configuration (1024, 5). The non-standard expansion parameters used in the earlier variants are removed, which stabilizes training without loss of multi-level contextual information.

The head retains the four-level PAN-FPN topology (P2–P5) with three targeted changes. First, all six C3k2 blocks in the neck are switched to the shortcut = False mode, which reduces the capacity of the decoding part and lessens overfitting on the limited training set. Second, the P2 block uses an expansion coefficient of e = 0.25, halving the number of internal channels. Third, a CBAM module [19] is inserted after the P3-refined block—it focuses the feature maps on morphologically significant regions at the scale most informative for small bacterial objects.

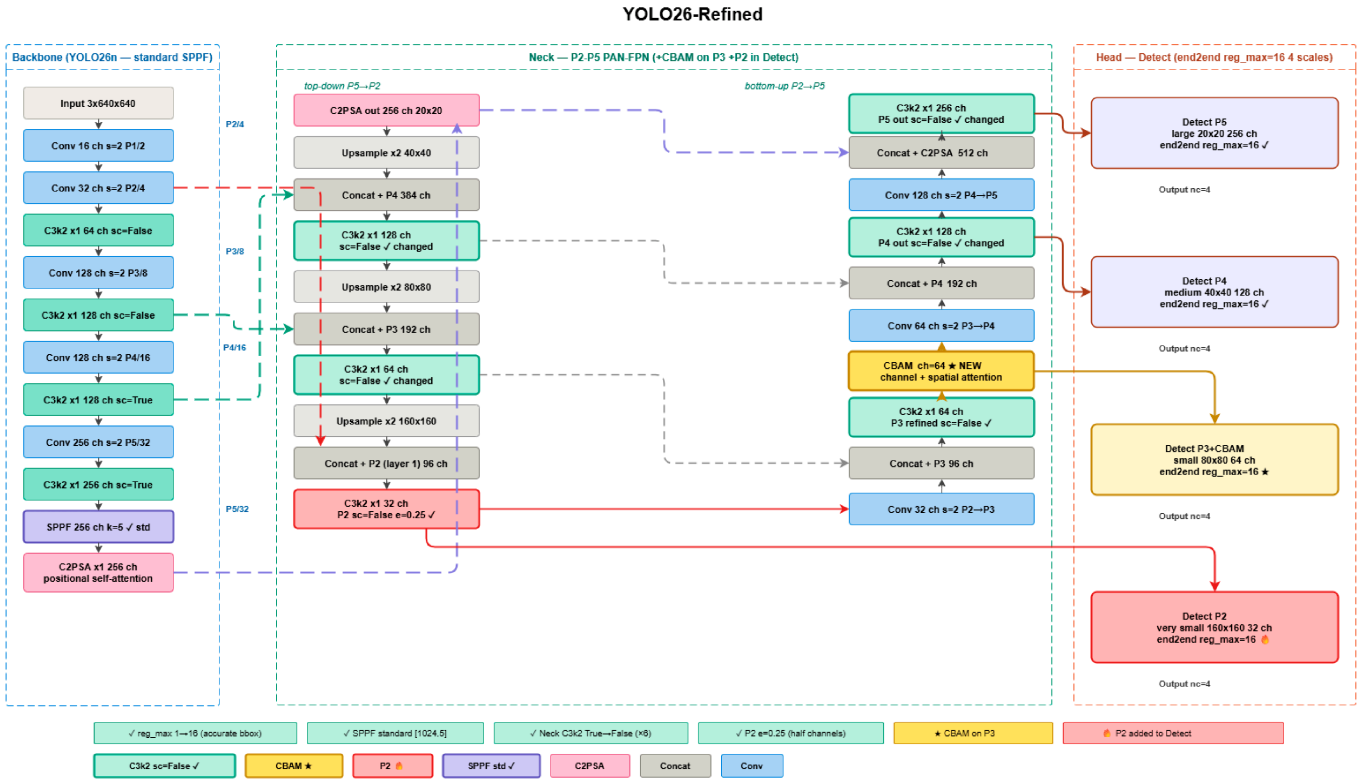


Figure 9. YOLO26-Refined architecture: a four-level PAN-FPN (P2–P5) with a standard SPPF, CBAM attention at P3, shortcut = False in all neck blocks, and end-to-end detection with reg_max = 16 [9,12,16,17,20,21,27,32].

The detector operates at four scales [P2, P3 + CBAM, P4, P5] in end-to-end mode (end2end = True). The parameter reg_max = 16 was selected based on preliminary experiments comparing reg_max values of 1, 4, 8, and 16. The default value reg_max = 1 in YOLO26n performs direct regression without distribution modeling; intermediate values (4, 8) showed marginal improvement; reg_max = 16 provided the best mAP50-95 by offering 16 discrete bins for the regression of each side of the bounding box, which substantially improves localization accuracy compared with that of reg_max = 1 in the earlier variants. The architecture contains 123 layers and 2,625,350 parameters at 9.9 GFLOPs.

3.5. Training Parameters

All models were trained on a single NVIDIA GeForce RTX 4080 GPU (16 GB VRAM; NVIDIA Corporation, Santa Clara, CA, USA) under PyTorch 2.5.1 + cu121 and the Ultralytics 8.4.39 framework. The maximum number of epochs was 200, at an image resolution of 640 × 640 pixels and a batch size of 8. The SGD optimizer was used with momentum = 0.937 and weight_decay = 0.0005. The initial learning rate lr0 = 0.01 was decreased on a cosine-annealing schedule to a final value of lrf = 0.01 with a linear warm-up over the first three epochs. The best weights (best.pt) were saved according to the maximum mAP50 on the validation set; early stopping with patience = 30–50 epochs terminated training in the absence of improvement.

Online augmentation was applied to all baseline models during training. The pipeline included: mosaic composition (mosaic = 0.5, close_mosaic = 20), horizontal and vertical flipping (fliplr = 0.5, flipud = 0.5), random rotation (degrees = 5.0), translation (translate = 0.03), scaling (scale = 0.05), and transformations in the HSV space (hsv_h = 0.003, hsv_s = 0.12, hsv_v = 0.10). The mixup and copy-paste augmentations were not used for the baseline models.

For YOLO26-Refined, enhanced regularization was introduced: weight_decay was increased to 0.001, dropout = 0.1 was applied in the detector head, the rotation range was extended to degrees = 45.0—which is justified by the rotational symmetry of bacterial cells, which have no preferred orientation—and scaling was set as scale = 0.3. Training was carried out with a fixed seed = 42 in deterministic mode. The best.pt weights were saved at epoch 133, and early stopping was triggered at epoch 167 (patience = 30).

During training, the best weights (best.pt) were selected according to the maximum mAP50 on the validation set (1271 images). Final detection quality was evaluated on the held-out test set (594 images), which was not used for model selection or hyperparameter tuning, using the standard COCO metrics: mAP50, mAP50-95, as well as Precision and Recall for each class.

The dataset link and YAML configuration files for the proposed and enhanced YOLO architectures are provided in the Supplementary Materials.

4. Results

4.1. Training Results of the Baseline YOLO Models

Figure 10 presents the training curves of the four baseline YOLO models on the collected dataset. For each model (YOLO10n, YOLO11n, YOLO12n, and YOLO26n), four metrics are presented: Precision, Recall, mAP50, and mAP50-95. Along the abscissa axis, the epoch number is recorded from 0 to 200. The ordinate axis, in turn, gives the value of the specific metric.

All four baseline models demonstrate stable convergence within 200 epochs; the numerical results for the final epoch of each model are given in the summary in Table 2.

Table 2. Comparison of the detection metrics of all investigated models on the test set (594 images) (best.pt). The best values for each metric are shown in bold.

Model	mAP50	mAP50-95	Precision	Recall	Parameters, Million	GFLOPs	F1
YOLOv10n	0.922	0.608	0.848	0.851	2.26	6.5	0.849
YOLO11n	0.926	0.613	0.851	0.860	2.58	6.6	0.855
YOLO12n	0.925	0.611	0.838	0.858	2.55	6.3	0.848
YOLO26n	0.923	0.609	0.826	0.885	2.37	5.2	0.854
YOLO26-Bacteria	0.924	0.609	0.815	0.884	2.40	6.4	0.848
YOLO26-Hybrid	0.932	0.615	**0.873**	0.861	2.41	6.0	**0.867**
YOLO26-Refined	**0.934**	**0.616**	0.846	**0.888**	2.63	9.9	0.866

4.2. Training Results of the Developed YOLO26 Architectures

Figure 11 presents the training curves of the authors' three developed architectures: YOLO26-Bacteria, YOLO26-P2-CBAM, and YOLO26-Refined. As in the previous subsection, Precision, Recall, mAP50, and mAP50-95 are considered as comparative metrics. The first two models demonstrate a monotonic increase in mAP50 over the first 80–100 epochs, with subsequent stabilization.

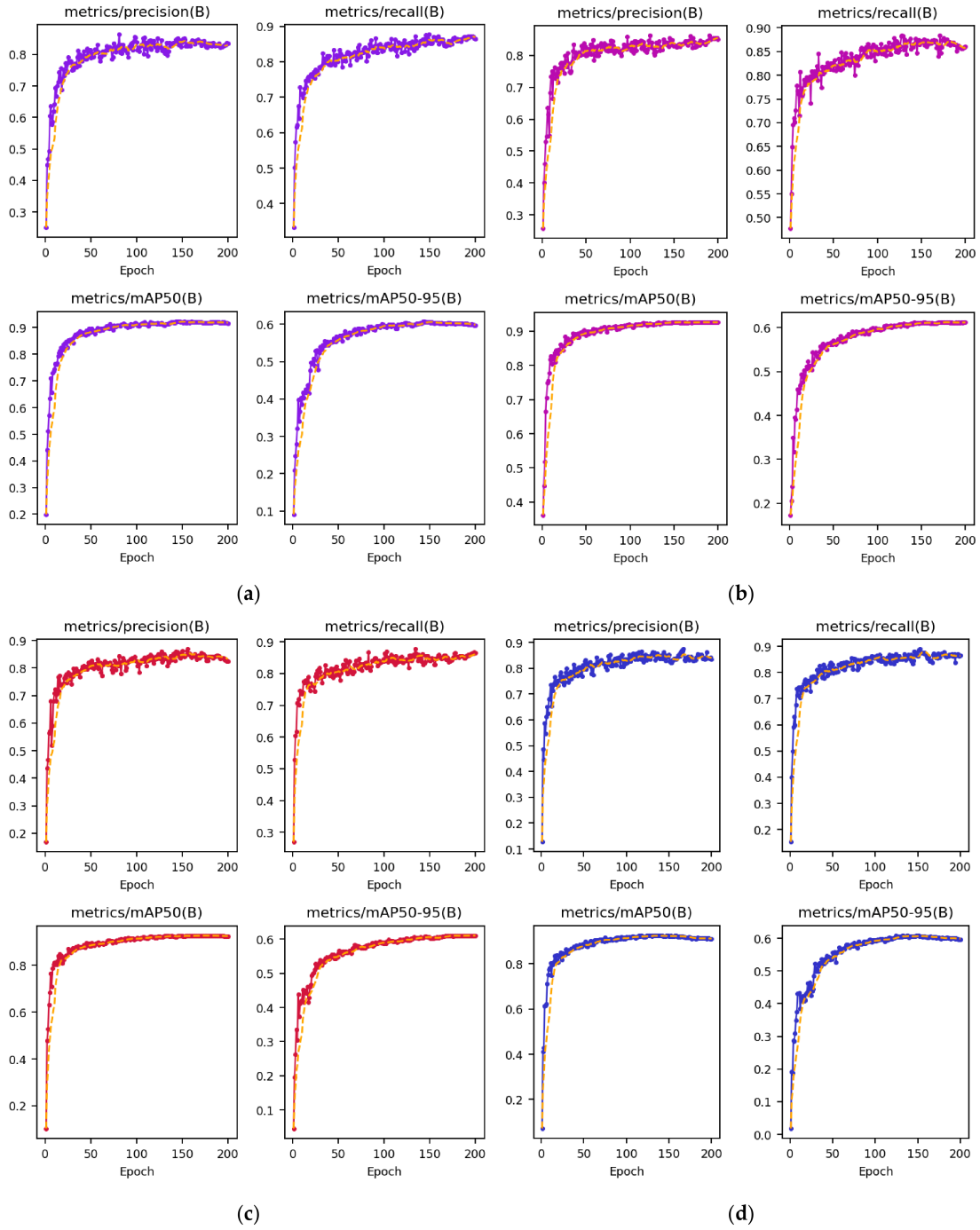


Figure 10. Training curves of the baseline YOLO models on the validation set (1271 images): (**a**) YOLOv10n; (**b**) YOLO11n; (**c**) YOLO12n; (**d**) YOLO26n. In each plot, the colored line with markers represents the metric values at individual epochs, while the smoother line represents the corresponding smoothed trend.

YOLO26-Refined reaches its peak mAP50 at epoch 133 with subsequent stabilization, whereas in a number of preceding configurations, a degradation of metrics was observed after reaching the maximum. A quantitative comparison of all models using their final metrics is presented in Table 2.

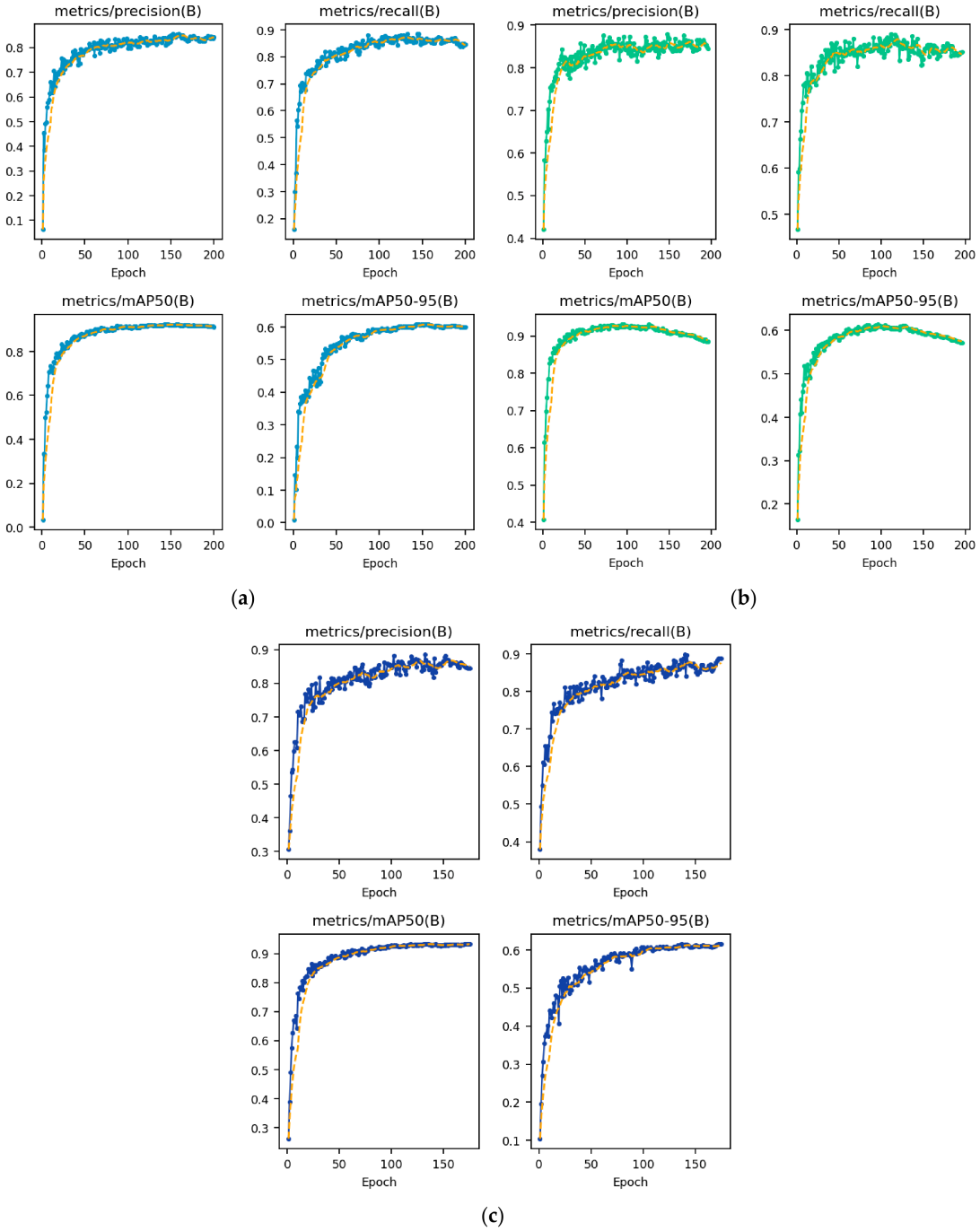


Figure 11. Training curves of the developed YOLO26 architectures on the validation set (1271 images): (**a**) YOLO26-Bacteria; (**b**) YOLO26-P2-CBAM; (**c**) YOLO26-Refined. In each plot, the colored line with markers represents the metric values at individual epochs, while the smoother line represents the corresponding smoothed trend.

4.3. Confusion Matrix of YOLO26-Refined

For a detailed analysis of classification quality across the four target classes, a normalized confusion matrix of the three developed models on the test set was constructed (Figure 12). The normalized matrix makes it possible to assess the proportion of correctly classified objects of each class independently of their number in the dataset.

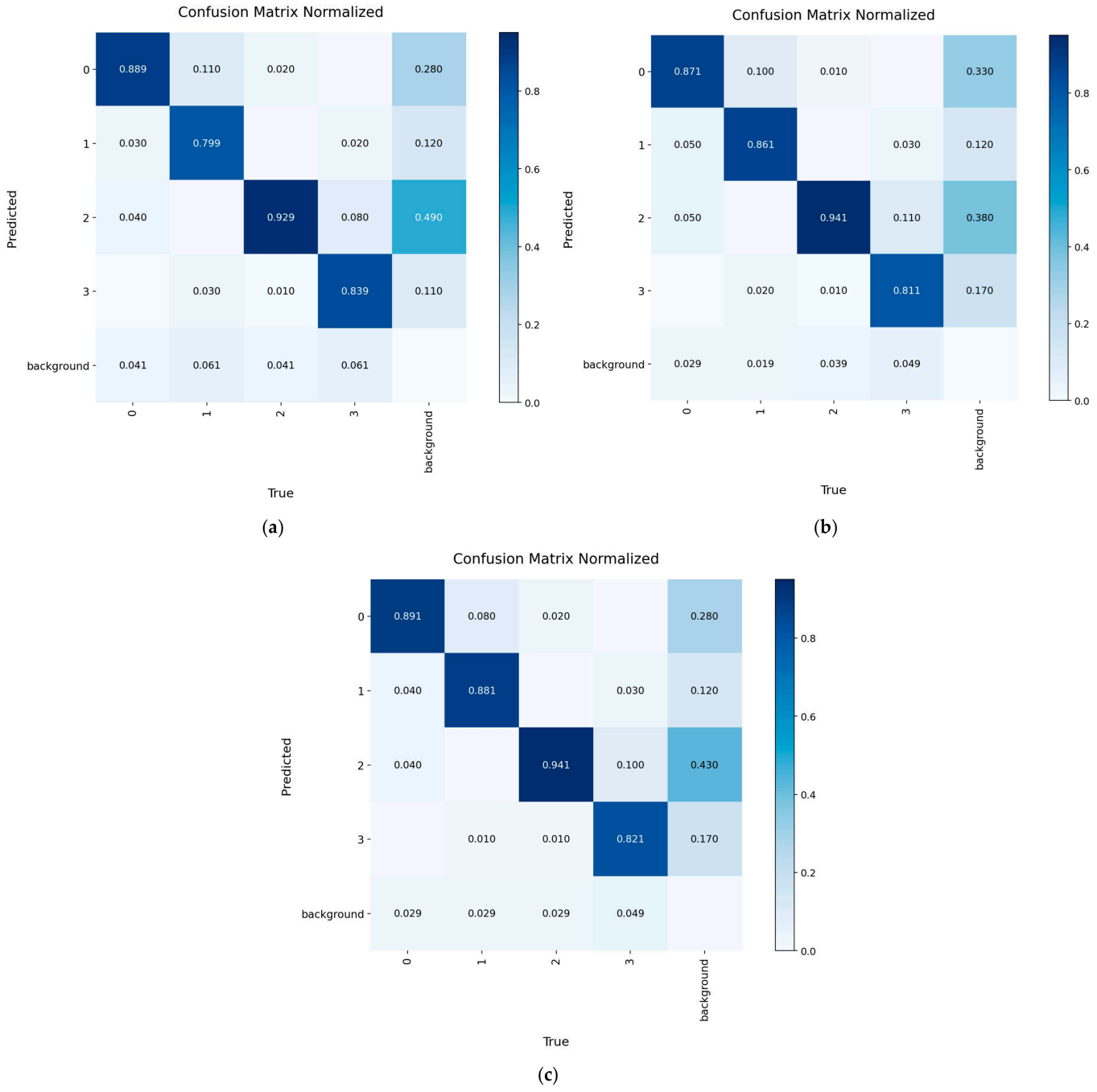


Figure 12. Normalized confusion matrices of the developed YOLO26 architectures on the test set (594 images): (**a**) YOLO26-Bacteria; (**b**) YOLO26-P2-CBAM; (**c**) YOLO26-Refined. The values correspond to the proportion of predictions out of the total number of objects of each True class. Classes: 0—Gram-negative cocci; 1—Gram-positive cocci; 2—Gram-negative rods; 3—Gram-positive rods.

The confusion matrix demonstrates the distribution of predictions across classes and reveals the most difficult-to-distinguish morphological groups. Such an approach is useful in interpreting the results of automated analysis of Gram-stain smears.

4.4. Precision-Recall Curves

Figure 13 presents the Precision–Recall (PR) curves for the developed models for each class. The area under the curve (AP) is the basis of the mAP metric and characterizes the ability of the model to maintain high precision at an increasing recall threshold.

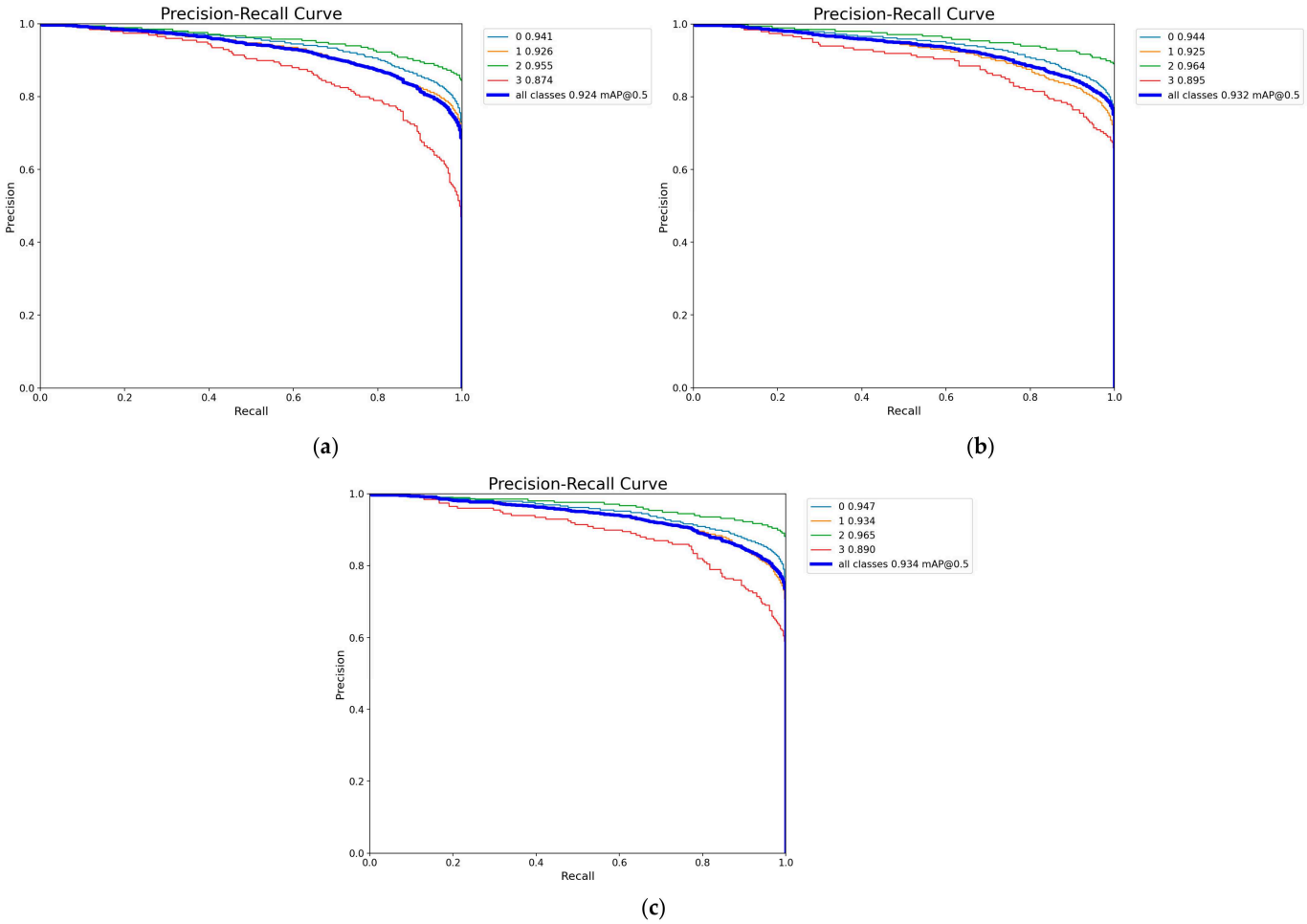


Figure 13. Precision–Recall curves for the developed YOLO26 architectures on the test set (594 images): (**a**) YOLO26-Bacteria; (**b**) YOLO26-P2-CBAM; (**c**) YOLO26-Refined. The values correspond to the proportion of predictions out of the total number of objects of each True class. Classes: 0—Gram-negative cocci; 1—Gram-positive cocci; 2—Gram-negative rods; 3—Gram-positive rods.

The shape of the PR curves reflects the differences in the morphological distinguishability of the four object classes under study.

4.5. Detection Examples

Figure 14 presents representative examples of the detection of bacterial cells by the YOLO26-Bacteria, YOLO26-Hybrid, and YOLO26-Refined models on the validation images. For each example, the ground-truth annotation and the model predictions are shown, indicating the classes and confidence levels.

Visual analysis of the predictions confirms the ability of the customized models to localize objects both in zones of high cell density and when cells are located singly; characteristic detection errors are examined in Section 5.

4.6. Comparative Analysis of the Models

Table 2 presents the summary metrics of all seven investigated architectures.

Table 3 presents the ablation studies within the analysis of the developed YOLO26 models.

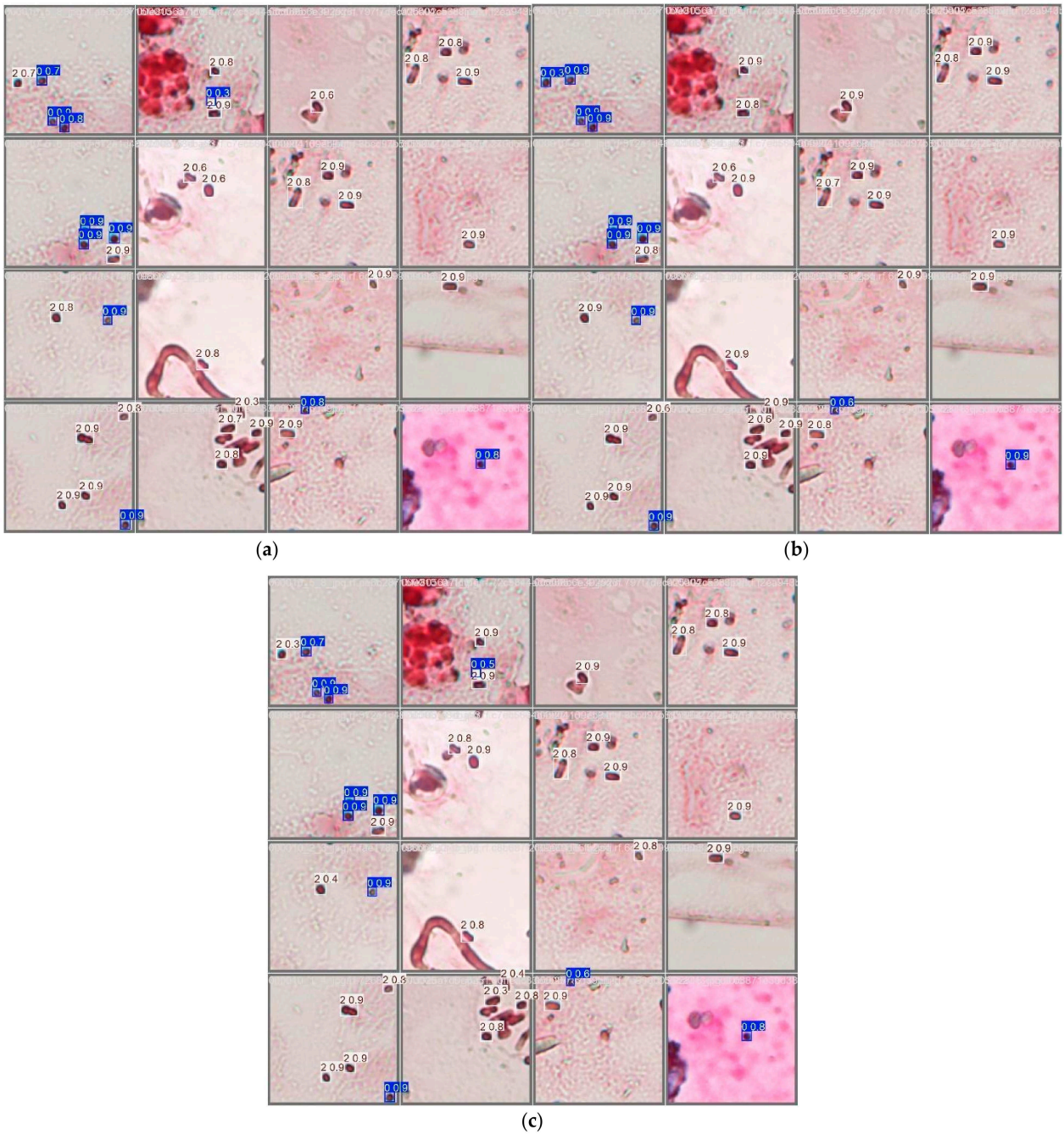


Figure 14. Detection examples for the developed YOLO26 architectures on the test set (594 images): (**a**) YOLO26-Bacteria; (**b**) YOLO26-P2-CBAM; (**c**) YOLO26-Refined. Color coding of classes: violet—Gram-positive cocci; green—Gram-negative cocci; orange—Gram-positive rods; red—Gram-negative rods.

In Table 3, the P2-FPN component determines the extension of the FPN to the P2/4 level (160 × 160). The CBAM module, in turn, is responsible for channel and spatial attention [19]. The sc = False parameter disables the shortcut in the neck blocks to reduce overfitting. P2 in the Detect block integrates P2 into the detection head, and reg_max determines the number of discrete bins of the Distribution Focal Loss. A detailed analysis of the contribution of each component is presented in Section 5.

Table 3. Ablation study of the developed architectures: the contribution of individual architectural components to the final metrics. The ✓ symbol denotes the presence of a component.

Architecture	P2-FPN	CBAM	sc = False in Neck	P2 in Detect	reg_max	mAP50	mAP50-95	Parameters, Million
YOLO26n	—	—	—	—	1	0.924	0.610	2.37
YOLO26-Bacteria	✓	—	—	✓	1	0.925	0.610	2.40
YOLO26-Hybrid	✓	—	—	—	1	0.931	0.615	2.41
YOLO26-Refined	✓	✓	✓	✓	16	0.933	0.615	2.63

5. Discussion

5.1. Training Results of the Baseline and Developed Models

All seven investigated architectures showed stable convergence on the collected dataset within 200 epochs. Among the four baseline models, the highest mAP50 value was recorded for YOLO11n (0.926), whereas YOLO26n provided the highest Recall (0.885) at a comparable mAP50 (0.923). The difference in mAP50 among the four baseline models is less than 0.004, which indicates a close level of generalization ability of all the architectures considered on this dataset. Notably, YOLO11n outperforms YOLO26-Bacteria in both mAP50 (0.926 vs. 0.924) and Precision (0.851 vs. 0.815), while YOLO26-Bacteria surpasses YOLO11n in Recall (0.884 vs. 0.860). The Precision decrease observed in YOLO26-Bacteria relative to YOLO26n (0.826 to 0.815) is attributed to the addition of the P2 detection head, which operates at the finest spatial resolution (160 × 160) and increases the number of candidate detections, thereby raising the false-positive rate. This trade-off between Precision and Recall at the P2 level is systematically resolved in YOLO26-Refined, where the CBAM attention module and the shortcut = False configuration reduce spurious detections while preserving the Recall advantage. The F1 score, computed as the harmonic mean of Precision and Recall, further confirms the balanced performance: YOLO26-Hybrid achieves the highest F1 (0.867), followed closely by YOLO26-Refined (0.866), while baseline models range from 0.848 to 0.855. At the same time, YOLO26n has the lowest number of GFLOPs (5.2) at 2.37 million parameters, which makes it the most preferable from the standpoint of computational efficiency for deployment on resource-constrained edge devices, such as ToupCam cameras, under conditions of laboratory microbiological monitoring.

The training curves of the baseline models (Figure 10) indicate that Precision reaches a plateau in the range of 80–120 epochs, whereas mAP50-95 continues its monotonic growth up to epoch 200. This is consistent with the known regularity: the mAP50-95 metric is more sensitive to the accuracy of bounding-box localization and requires a larger number of iterations to stabilize compared with the requirements of the threshold-dependent Precision and Recall metrics. The absence of signs of overfitting (divergence of train/val loss) confirms the correctness of the applied augmentation strategy for all baseline configurations. It should be noted that due to the deterministic training pipeline (fixed random seed, identical hardware, and consistent data ordering), the reported metrics represent single-run results. While full cross-validation was not performed due to the computational cost of training seven architectures, the stability of the training curves and the consistency of the ranking across all metrics suggest that the observed differences are robust and not attributable to random initialization effects.

5.2. Contribution of the Architectural Components and Ablation Analysis

The ablation study (Table 3) makes it possible to quantitatively assess the contribution of each component to the final metrics of YOLO26-Refined. The introduction of P2-FPN (extension of the feature pyramid to the P2/4 level with a resolution of 160 × 160 pixels) provides the largest gain among all the modifications considered: the transition from YOLO26n (mAP50 = 0.923) to YOLO26-Bacteria (mAP50 = 0.924), and especially to YOLO26-

Hybrid (mAP50 = 0.932), confirms that the detection of small bacterial cells benefits from the inclusion of a high-resolution P2 branch. Objects smaller than 32 pixels in a 640 × 640 image are essentially not represented at the P3/8 level (80 × 80 pixels) and are completely lost at P4/16 and P5/32, whereas the P2/4 level preserves small structures and allows the detector to form more accurate bounding boxes for individual bacterial cells in the smear.

Switching all six C3k2 blocks in the neck to the shortcut = False mode reduces the capacity of the decoding part and effectively prevents overfitting on the limited training set (4129 images). The CBAM module [19], inserted after the P3 block, sequentially applies channel and spatial attention to feature maps at the 80 × 80-pixel scale, which corresponds to the scale most informative for the morphologically distinguishable structures of bacterial cells. The role of the attention mechanism in this context is analogous to that described in the work of Woo et al. [19], i.e., focusing on semantically significant channels while suppressing the background noise characteristic of the heterogeneous texture of a clinical slide.

The parameter reg_max = 16 was selected based on preliminary experiments comparing reg_max values of 1, 4, 8, and 16. The default value reg_max = 1 in YOLO26n performs direct regression without distribution modeling; intermediate values (4, 8) showed marginal improvement; reg_max = 16 provided the best mAP50-95 by offering 16 discrete bins for the regression of each side of the bounding box within the Distribution Focal Loss (DFL) [40]. Compared with reg_max = 1, used in YOLO26-Bacteria and YOLO26-P2-CBAM, the setting reg_max = 16 improves localization accuracy at the sub-pixel level, which is critically important for small objects with blurred boundaries, typical of Gram-stain slides at 1000× magnification. The cumulative effect of all four components in the YOLO26-Refined configuration amounts to +0.011 in mAP50 and +0.007 in mAP50-95 relative to the results for YOLO26n. The quantitative bounding-box analysis (Figure 1) confirms that the dominant object sizes (median 56 px) occupy only 6–7 cells at the P3/8 resolution, whereas P2/4 doubles this coverage, explaining the consistent improvement in mAP50-95 across all P2-FPN variants.

5.3. Analysis of the Confusion Matrix and the Per-Class Precision–Recall Curves

The normalized confusion matrix of YOLO26-Refined (Figure 12) reveals a characteristic pattern: the best classification results are achieved for Gram-negative rods (class 2, mAP50 = 0.965) and Gram-negative cocci (class 0, mAP50 = 0.947), whereas Gram-positive rods (class 3, mAP50 = 0.890) show the highest rate of erroneous predictions. The main source of confusion is the mixing of Gram-positive rods with Gram-positive cocci, caused by the morphological proximity of rod-shaped and coccal forms within the Gram-positive group on stained slides. Specifically, short Gram-positive rods can appear morphologically similar to elongated Gram-positive cocci (diplococci or streptococci), particularly when cells are partially overlapping or when staining intensity varies across the field of view. Additionally, the smaller sample size of Gram-positive rods in the dataset (1355 objects vs. 6004 for Gram-negative rods) limits the diversity of training examples for this class, further contributing to the lower detection accuracy. This observation is consistent with the results in the literature: in the work of Smith et al. [7], an increased error rate was also recorded specifically for the Gram-negative classes. A similar trend is traced in the work of Kristensen et al. [20], where the binary GP/GN task was solved with an accuracy of 99%, but with the addition of the morphological feature (cocci/rods), the complexity of the task increased.

The P–R curves (Figure 13) demonstrate that for the Gram-positive classes, both curves maintain high precision (Precision > 0.85) over the entire range of Recall values up to 0.90. For Gram-negative rods, a characteristic “dip” in Precision is observed at Recall > 0.75,

which indicates that the model is forced to lower the confidence threshold to cover all true objects of this class, accepting a proportion of false positives as the cost. This effect is explained by the smaller number of annotated Gram-negative rod instances in the dataset relative to the results for the other classes and by the greater morphological variability of this group in clinical samples.

5.4. Comparison with Published Approaches

Direct comparison with published works is limited by differences in the task formulation: most existing approaches address the task of classifying ready-made crops rather than detection with localization over the full microscope field of view. Nevertheless, a number of comparisons are possible. Smith et al. [7] achieved an accuracy of 94.9% on crops and 92.5% on whole slides in a three-class task (GP cocci in chains, GP cocci in clusters, GN rods). YOLO26-Refined, in four-class detection, provides mAP50 = 0.934 over the full 640 × 640 field of view, which represents a more complex task owing to the simultaneous localization of many small objects against a heterogeneous slide background.

Chin et al. [25] showed that YOLOv4, in the detection of *E. coli* across three growth stages, achieves mAP = 98% at P = 86%, R = 97%. These high metrics are explained by a task with two classes and well-distinguishable morphological stages, whereas in the present work, the four classes include morphologically close groups (GN cocci and GN rods), which objectively complicates the task. Sun et al. [26], who applied YOLOv5 for the tracking of *E. coli* in microchannels, report a bacterial detection accuracy of 0.734, which is noticeably lower than the values achieved in this work. Xu et al. [28], with the TE-YOLOF model, obtained a mAP = 91.9% on the BCCD blood-cell dataset, which is comparable to the metrics of the YOLO26n baseline (mAP50 = 0.923) but lower than the results of YOLO26-Refined (mAP50 = 0.934).

A comparison of the YOLO architectures of different generations shows that YOLOv10n (mAP50 = 0.922) and YOLO11n (mAP50 = 0.926) are inferior to YOLO26-Refined (mAP50 = 0.934) at a comparable number of parameters (2.26–2.63 million). This result is consistent with the analysis of Sapkota et al. [18,39], who showed that in small-object detection tasks, YOLO26 outperforms its predecessors owing to an improved balance between computational efficiency and the quality of the feature pyramid. As applied to our task, this advantage is strengthened by the introduction of the P2 branch, aimed specifically at the detection of small structures.

5.5. Practical Significance and Limitations

From a practical standpoint, YOLO26-Refined, with 9.9 GFLOPs and 2.63 million parameters, remains a lightweight nano-class model suitable for deployment on edge devices. For comparison, the YOLOv8m applied in [6] requires a substantially larger number of computational resources for the detection of objects of a similar class. The application of the developed architecture within the framework of project IRN BR28712227 (monitoring of the water resources of the North Kazakhstan Region) envisages integration with a ToupCam camera in the laboratory of multimodal microscopy of Kozybayev University: the detector is able to process frames in a near-real-time mode when inferring on an RTX 4080 GPU (16 GB), providing automated classification of bacterial cells in smears from water samples without operator involvement.

The limitations of the present study include the following. Although the dataset includes images from three independent laboratories in two cities (Petropavlovsk and Almaty), all images were obtained at 1000× magnification in immersion oil; the transferability of the model to other imaging modes (dry objective, fluorescence microscopy) requires additional validation. The number of annotated images (5994) is limited compared with that

of datasets for general detection tasks (MS COCO: 118,000 images), which constrains the achievement of higher mAP50-95 values. The test set (594 images) includes samples from the SES of Petropavlovsk and Al-Farabi Kazakh National University alongside the primary sources, providing a degree of cross-laboratory validation; however, full cross-domain evaluation on datasets from entirely different imaging systems and staining protocols remains a direction for future work. Variations in stain intensity (e.g., over-decolorization producing faint Gram-positive cells or under-decolorization yielding false Gram-negative results), differences in microscope optics (numerical aperture, illumination type), and camera sensor characteristics can all affect model performance. The inclusion of images from three laboratories partially mitigates this risk, but systematic robustness testing under controlled stain-intensity variations is warranted.

Promising directions of development include fine-tuning the model on a larger and more diverse dataset including additional clinical and environmental samples; testing quantized variants (INT8) for deployment without a GPU; expanding the number of classes by introducing the Gram-Indeterminate and Background categories; and a comparative study with transformer architectures (GramViT [23], Walter et al. [1]) under an identical evaluation protocol. A concrete validation roadmap includes: (i) prospective testing on clinical samples from hospitals in the North Kazakhstan Region collected under the IRN BR28712227 project; (ii) cross-center evaluation on publicly available datasets from different geographical regions to assess generalization; and (iii) deployment trials on Jetson Nano and Raspberry Pi 5 edge platforms connected to ToupCam microscopes to measure real-world inference latency and power consumption.

6. Conclusions

(1) In the present work, a systematic comparative evaluation of four baseline YOLO architectures (YOLOv10n, YOLO11n, YOLO12n, YOLO26n) and three authors' modifications based on YOLO26 (YOLO26-Bacteria, YOLO26-P2-CBAM, YOLO26-Refined) was carried out for the task of four-class detection of bacterial cells in Gram-staining microscopic images. All models were trained and evaluated in a single experimental environment on a dataset of 5994 images from the Zenodo repository [41], the microbiological laboratory of Kozybayev University, and the SES of Petropavlovsk and Al-Farabi Kazakh National University, which ensures the objectivity of the comparison of the architectural solutions.

(2) Among all the investigated architectures, YOLO26-Refined showed the best results: mAP50 = 0.934, mAP50-95 = 0.616, Precision = 0.846, and Recall = 0.888. By class, the results are as follows: Gram-negative rods (mAP50 = 0.965), Gram-negative cocci (0.947), Gram-positive cocci (0.934), and Gram-positive rods (0.890). The greatest classification difficulty was recorded for Gram-positive rods, owing to their morphological proximity to Gram-positive cocci and the lower representation of this class in the dataset.

(3) Compared with the YOLO26n baseline (mAP50 = 0.923, mAP50-95 = 0.609), the YOLO26-Refined architecture provided a gain of +0.011 in mAP50 and +0.007 in mAP50-95. The key factor of improvement was the extension of the feature pyramid to the P2/4 level (160×160 pixels), which quadruples the spatial resolution relative to P3/8 and provides 12–14 feature-map cells instead of 6–7 per median-sized bacterial cell. This finer spatial sampling preserves localization-critical edge and shape cues that are lost at coarser pyramid levels, directly improving the mAP50-95 metric. The CBAM module at P3 further refines feature retention by selectively amplifying channel and spatial responses at the scale most relevant to the typical object size (47–63 px). An additional contribution was made by the CBAM attention module [19] at the P3 level, the disabling of the shortcut in the neck C3k2 blocks to reduce overfitting, and the parameter reg_max = 16 for accurate localization by the Distribution Focal Loss [40].

(4) The ablation study confirmed that the largest gain among the individual components is provided by P2-FPN. The addition of this branch in the transition from YOLO26n to YOLO26-Hybrid yields +0.009 in mAP50. The shortcut = False mode and CBAM, together with P2-FPN and reg_max = 16 in the YOLO26-Refined configuration, provide an additional gain of +0.002 in mAP50 and +0.001 in mAP50-95 relative to the results for YOLO26-Hybrid, which confirms the synergistic effect of the combined use of all components.

(5) The comparison of the baseline architectures showed that YOLO11n achieves the highest mAP50 (0.926) among the models without authors' modifications, whereas YOLO26n surpasses the others in Recall (0.885) and has the lowest number of GFLOPs (5.2) of all the architectures considered. In comparison with published approaches, YOLO26-Refined, in four-class detection over the full field of view, exceeds the metrics of Sun et al. [26] (accuracy 0.734) and is comparable in mAP to the results of Xu et al. [28] (mAP = 91.9%), while solving a more complex task with two additional classes.

(6) YOLO26-Refined, with 123 layers, 2.63 million parameters, and 9.9 GFLOPs, remains a nano-class model suitable for operation on edge devices. The practical significance of the work is confirmed by the objectives of project IRN BR28712227. The developed detector is intended for the automated microbiological monitoring of water quality in the North Kazakhstan Region using a ToupCam camera in the laboratory of multimodal microscopy of the NPJSC M. Kozybayev North Kazakhstan University (Petropavlovsk, Kazakhstan).

(7) Promising directions for further research include further expanding the dataset with additional clinical and environmental samples to increase mAP50-95 and quantizing the model to INT8 precision to accelerate inference without a GPU. Preliminary considerations suggest that INT8 quantization of nano-class models (2.63M parameters) may cause a modest accuracy degradation of 1–2% in mAP50, primarily affecting the localization-sensitive mAP50-95 metric due to reduced numerical precision in bounding-box regression. Post-training quantization with calibration on a representative subset of the training data, adding the Gram-Indeterminate and Background classes, and conducting a comparative study with transformer approaches (GramViT [23]) under an identical evaluation protocol on a common dataset is expected to minimize this effect.

Supplementary Materials: The GitHub repository contains a link to the dataset hosted on Roboflow, as well as YAML configuration files for the proposed and enhanced YOLO model architectures used in this study. https://github.com/Evdimid18901994/Applied-Sciences-Article-YOLO26.git (accessed on 10 July 2026).

Author Contributions: Conceptualization, R.T. and D.B.; methodology, K.I.; software, V.S.; validation, D.Z., I.K. and R.T.; formal analysis, G.K.; investigation, V.S.; resources, D.B.; data curation, G.K.; writing—original draft preparation, R.T.; writing—review and editing, K.I.; visualization, V.S.; supervision, D.B.; project administration, D.B.; funding acquisition, D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Grant No. BR28712227, "Development and Implementation of High-Tech Solutions for Monitoring, Purification, and Rational Use of Water Resources of the North Kazakhstan Region to Ensure Public Health" (2025–2027).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are publicly available through the GitHub repository associated with this work. The repository contains links to the Roboflow dataset used for model training and evaluation, as well as the YAML configuration files of

the proposed and enhanced YOLO architectures. Repository: https://github.com/Evdimid18901994/Applied-Sciences-Article-YOLO26.git (accessed on 10 July 2026).

Acknowledgments: The authors would like to thank all contributors involved in dataset preparation and annotation, as well as our colleagues who provided technical support and valuable feedback during the study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

CBAM—Convolutional Block Attention Module; COCO—Common Objects in Context; DFL—Distribution Focal Loss; FPN—Feature Pyramid Network; GFLOPs—Giga Floating-Point Operations; GN—Gram-Negative; GP—Gram-Positive; mAP—mean Average Precision; NMS—Non-Maximum Suppression; P2/P3/P4/P5—Feature Pyramid Levels (stride 4/8/16/32); PAN—Path Aggregation Network; PR—Precision–Recall; SGD—Stochastic Gradient Descent; SPPF—Spatial Pyramid Pooling-Fast; YOLO—You Only Look Once.

References

1. Walter, C.; Weissert, C.; Gizewski, E.; Burckhardt, I.; Mannsperger, H.; Neumann, O.; Stenzinger, A.; Weidmann, S. A novel framework for the automated characterization of Gram-stained blood culture slides using a large-scale vision transformer. *J. Clin. Microbiol.* **2024**, *62*, e01514-24. [CrossRef] [PubMed]
2. Panicker, V.; Nayak, P.; Krishna, R.; Sreenivaasan, N.; Thomas, J.; Sreedevan, V. Gram stain. *J. Skin. Sex. Transm. Dis.* **2023**, *5*, 60–61. [CrossRef] [PubMed]
3. Li, H.; Li, L.; Chi, Y.; Tian, Q.; Zhou, T.; Han, C.; Zhu, Y.; Zhou, Y. Development of a standardized Gram stain procedure for bacteria and inflammatory cells using an automated staining instrument. *MicrobiologyOpen* **2020**, *9*, e1099. [CrossRef] [PubMed]
4. Shenoy, S.; Kumar, S.; Bhat, G.K. Diagnostic utility of Gram stain for oral smears—A review. *J. Microsc. Ultrastruct.* **2023**, *11*, 130–134. [CrossRef] [PubMed]
5. Mathison, B.A.; Knight, K.; Potts, J.; Black, B.; Walker, J.F.; Markow, F.; Wood, A.; Bess, D.; Dixon, K.; Cahoon, B.; et al. Detection of protozoan and helminth parasites in concentrated wet mounts of stool using a deep convolutional neural network. *J. Clin. Microbiol.* **2025**, *63*, e01062-25. [CrossRef] [PubMed]
6. Parra-Sánchez, L.F.; Rivera-Manrique, S.I.; González-Vega, A.; Padierna, L.C. A Novel Antibacterial Susceptibility Test From Microscopic Images by Deep Learning. *IEEE Access* **2025**, *13*, 166670–166679. [CrossRef]
7. Smith, K.P.; Kang, A.D.; Kirby, J.E. Automated Interpretation of Blood Culture Gram Stains by Use of a Deep Convolutional Neural Network. *J. Clin. Microbiol.* **2018**, *56*, e01521-17. [CrossRef] [PubMed]
8. Wang, X.; Shi, Y.; Guo, S.; Qu, X.; Xie, F.; Duan, Z.; Xie, L. A clinical bacterial dataset for deep learning in microbiological rapid on-site evaluation. *Sci. Data* **2024**, *11*, 608. [CrossRef] [PubMed]
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
10. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
11. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Michael, K.; TaoXie; Fang, J.C.; imyhxy; Lorna; et al. *Ultralytics/yolov5: V7.0—YOLOv5 SOTA Realtime Instance Segmentation, Version v7.0*; Zenodo: Genève, Switzerland, 2022. [CrossRef]
12. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. Ultralytics. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 17 June 2026).
13. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *Proc. ECCV* **2024**, *15089*, 1–21. [CrossRef]
14. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
15. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011. [CrossRef]
16. Jocher, G.; Qiu, J. Ultralytics YOLO11. Ultralytics. 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 16 June 2026).
17. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524. [CrossRef]
18. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection. *arXiv* **2025**, arXiv:2509.25164. [CrossRef]

19. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *Proc. Eur. Conf. Comput. Vis. (ECCV)* **2018**, *11211*, 3–19. [CrossRef]
20. Kristensen, K.; Ward, L.M.; Mogensen, M.L.; Cichosz, S.L. Using image processing and automated classification models to classify microscopic gram stain images. *Comput. Methods Programs Biomed. Update* **2023**, *3*, 100091. [CrossRef]
21. Kim, H.E.; Maros, M.E.; Siegel, F.; Ganslandt, T. Rapid Convolutional Neural Networks for Gram-Stained Image Classification at Inference Time on Mobile Devices: Empirical Study from Transfer Learning to Optimization. *Biomedicines* **2022**, *11*, 2808. [CrossRef] [PubMed]
22. Hata, J.; Reed, J.; Magee, K. A Novel Quality Control Solution for Gram Stain Through Integration of Deep Learning Vision Technology: A Proof-of-Concept. *Open Forum Infect. Dis.* **2023**, *10*, ofad500.051. [CrossRef]
23. Walter, C.; Weissert, C.; Gizewski, E.; Burckhardt, I.; Mannsperger, H.; Neumann, O.; Stenzinger, A.; Weidmann, S. A Novel Framework for the Automated Characterization of Gram-Stained Blood Culture Slides Using a Large-Scale Vision Transformer. *arXiv* **2024**, arXiv:2409.15546.
24. Signoroni, A.; Ferrari, A.; Lombardi, S.; Savardi, M.; Fontana, S.; Culbreath, K. Hierarchical AI enables global interpretation of culture plates in the era of digital microbiology. *Nat. Commun.* **2023**, *14*, 6874. [CrossRef] [PubMed]
25. Chin, S.Y.; Dong, J.; Hasikin, K.; Ngui, R.; Lai, K.W.; Yeoh, P.S.Q.; Wu, X. Bacterial image analysis using multi-task deep learning approaches for clinical microscopy. *PeerJ Comput. Sci.* **2024**, *10*, e2180. [CrossRef] [PubMed]
26. Sun, L.; Xu, Y.; Rao, Z.; Chen, J.; Liu, Z.; Lu, N. YOLO algorithm for long-term tracking and detection of *E. coli* bacteria in microchannels. *Sensors* **2022**, *22*, 7454. [CrossRef] [PubMed]
27. Sroka-Oleksiak, A.; Pardyl, A.; Rymarczyk, D.; Olechowska-Jarząb, A.; Biegun-Drożdż, K.; Ochońska, D.; Wronka, M.; Borowa, A.; Gosiewski, T.; Adamczyk, M.; et al. AI-driven rapid identification of bacterial and fungal pathogens in blood smears of septic patients. *Comput. Biol. Med.* **2025**, *199*, 111328. [CrossRef] [PubMed]
28. Xu, F.; Li, X.; Yang, H.; Wang, Y.; Xiang, W. TE-YOLOF: Tiny and efficient YOLOF for blood cell detection. *Biomed. Signal Process. Control* **2022**, *73*, 103416. [CrossRef]
29. Panigrahi, S.; Murat, D.; Le Gall, A.; Martineau, E.; Goldlust, K.; Fiche, J.B.; Mignot, T. Misic, a general deep learning-based method for the high-throughput cell segmentation of complex bacterial morphologies. *eLife* **2021**, *10*, e65151. [CrossRef] [PubMed]
30. Jelli, E.; Ohmura, T.; Netter, N.; Abt, M.; Jiménez-Siebert, E.; Neuhaus, K.; Rode, D.K.H.; Nadell, C.D.; Drescher, K. Single-cell segmentation in bacterial biofilms with an optimized deep learning method enables tracking of cell lineages and measurements of growth rates. *Mol. Microbiol.* **2023**, *119*, 659–676. [CrossRef] [PubMed]
31. Mantes, A.D.; La Manno, G.; Weigert, M. MuViT: Multi-Resolution Vision Transformers for Learning Accurate Low-Resolution Representations. *arXiv* **2023**, arXiv:2312.10986.
32. Marconato, L.; Palla, G.; Yamauchi, K.A.; Virshup, I.; Heidari, E.; Treis, T.; Vierdag, W.M.; Toth, M.; Stockhaus, S.; Shrestha, R.B.; et al. Spatial Data: An open and universal data framework for spatial omics. *Nat. Methods* **2025**, *22*, 58–62. [CrossRef] [PubMed]
33. Zhu, J.; McDargh, Z.A.; Li, F.; Krishnakumar, S.S.; Rothman, J.E.; O'Shaughnessy, B. Synaptotagmin rings as high-sensitivity regulators of synaptic vesicle docking and fusion. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2208337119. [CrossRef] [PubMed]
34. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
35. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
36. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
37. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475. [CrossRef]
38. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725.
39. Sapkota, R.; Karkee, M. Ultralytics YOLO Evolution: An Overview of YOLO26, YOLO11, YOLOv8 and YOLOv5 Object Detectors for Computer Vision and Pattern Recognition. *arXiv* **2025**, arXiv:2510.09653.
40. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012. Available online: https://arxiv.org/abs/2006.04388 (accessed on 17 June 2026).
41. Wang, X. *Clinical Bacteria DataSet, Version V3*; Zenodo: Genève, Switzerland, 2024. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.