Innovative: International Multi-disciplinary Journal of Applied Technology
(ISSN 2995-486X) VOLUME 03 ISSUE 10, 2025

LIGHTWEIGHT YET PRECISE: INCEPTION-RESNET-A BACKBONE FOR YOLO-BASED WCE POLYP DETECTION

[+] *Saydirasulov S.N.*[1], Mukhamadiyev A.N.[1], Turimov D.M.[1], *Kilichev D.*[2]

[1] Department of IT Convergence Engineering Gachon University, Gyeongi-do, South Korea
[+] saydirasulov@gacho.ac.kr, mukhamadiyev@gacho.ac.kr

[2] Samarkand Branch of Tashkent University of Economics, Uzbekistan

Abstract:

Wireless capsule endoscopy (WCE) produces long, variable-quality video streams in which early and reliable polyp detection is critical. We present YOLO-InceptionResNet-A, a lightweight object detector that replaces the standard YOLOv4-tiny backbone with an Inception-ResNet-A block to enrich multi-scale feature representation while preserving real-time efficiency. The proposed pipeline operates in two stages: (i) a frame-level screening classifier to filter normal/abnormal images, and (ii) the detector for precise polyp localization. To respect clinical color sensitivity, we adopt conservative, clinically aware augmentation (brightness and mild hue jitter), alongside standard normalization. We evaluate on the Kvasir family of WCE images using patient-level splits and report object-detection metrics (mAP@0.5, mAP@[.5:.95], precision/recall/F1, and IoU), frame-level classification metrics (AUROC, sensitivity, specificity), and throughput on a single RTX 3090 GPU. Across benchmarks, our backbone swap consistently improves detection mAP and recall over YOLOv3, YOLOv4, and YOLOv4-tiny baselines, while maintaining low latency suitable for real-time review. Ablation studies isolate the contributions of the Inception-ResNet-A backbone and the augmentation policy, demonstrating that richer multi-scale features are the primary driver of the gains. We discuss limitations related to dataset size and domain shift, and outline external validation on additional WCE datasets as future work. These results indicate that targeted backbone re-architecture can deliver lightweight yet precise WCE polyp detection without sacrificing speed—an attractive trade-off for clinical deployment.

Keywords: Wireless Capsule Endoscopy (WCE), Polyp Detection, Object Detection, YOLO/YOLOv4-tiny, Inception-ResNet-A, Lightweight Backbone, Real-Time Detection, Medical Image Analysis, Gastrointestinal Endoscopy, Clinical AI.

1. Introduction

Wireless capsule endoscopy (WCE) enables non-invasive visualization of the gastrointestinal (GI) tract and has become integral to screening and surveillance workflows, particularly for early identification of colorectal lesions such as polyps [1,2]. However, WCE produces hours of video per patient, with variable illumination, specular highlights, motion blur, bubbles, and debris—factors that increase reader fatigue and the risk of missed findings in routine practice [3,4]. Automated computer-aided detection (CADe) and diagnosis (CADx) systems can mitigate this burden by prioritizing suspicious frames and localizing lesions for review, provided they meet stringent accuracy and latency requirements compatible with clinical use [5].

Deep learning has rapidly advanced GI imaging across segmentation, classification, and detection tasks. Encoder–decoder architectures (U-Net and its derivatives) dominate segmentation of mucosal structures and polyp boundaries [6–8], while convolutional neural networks (CNNs) such as ResNet and Inception-ResNet families remain strong baselines for frame-level abnormality classification [9,10]. For real-time localization, one-stage detectors—particularly the YOLO family—offer an attractive speed–accuracy trade-off by formulating object detection as a single regression problem [11–13]. YOLOv4 and the compact YOLOv4-tiny variant extend this paradigm to resource-constrained settings, but their lightweight backbones can under-represent fine, multi-scale textural cues that are critical for subtle, flat, or small polyps in WCE [12,14].

Backbone capacity and feature aggregation are central to detector performance. Inception-ResNet blocks combine residual learning with multi-branch receptive fields, enhancing the network's ability to capture scale diversity without incurring prohibitive computational cost [10,15]. We hypothesize that selectively upgrading the YOLOv4-tiny backbone with an Inception-ResNet-A block can enrich representational power where it matters most—early features and cross-scale fusion—while preserving the real-time throughput that makes tiny detectors clinically appealing.

Datasets for WCE vary in modality (image vs video), annotation granularity (classification labels, bounding boxes, or masks), and acquisition conditions. Public resources such as the Kvasir family provide standardized benchmarks and facilitate reproducible comparisons under patient-level splits that prevent identity leakage across train/validation/test partitions [16–18]. Given the color sensitivity of GI mucosa, augmentation policies must be conservative to avoid unrealistic chromatic shifts; small-magnitude brightness and hue jitter, combined with normalization, have been recommended to improve robustness without compromising clinical fidelity [19,20].

This work. We introduce YOLO-InceptionResNet-A, a lightweight detector for WCE polyp localization that replaces the default YOLOv4-tiny backbone module with an Inception-ResNet-A block. Our pipeline comprises: (i) a frame-level screening classifier to filter normal vs abnormal frames, and (ii) the detector for precise localization. We evaluate on Kvasir-style benchmarks using patient-level splits and report object-detection metrics—mAP@0.5, mAP@[.5:.95], precision, recall, and IoU—alongside throughput on a single RTX-class GPU. Ablation studies isolate the contributions of the backbone swap and augmentation policy. Our results show consistent gains in mAP and recall over YOLOv3/YOLOv4 baselines at comparable latency, supporting the premise that targeted backbone re-architecting can yield lightweight yet precise WCE detection suitable for clinical triage and review [11–18].

2. RELATED WORK

2.1 Computer-aided analysis in GI endoscopy

Classical CADe/CADx systems for GI endoscopy span segmentation, classification, and detection. Encoder–decoder models such as U-Net and its descendants (U-Net++, ResUNet++) dominate pixel-level segmentation of mucosal boundaries and polyps, improving delineation of small and flat lesions through dense skip connections and residual/attention modules [1–4]. For frame-level abnormality screening, high-capacity CNNs (e.g., ResNet-50/101, Inception-ResNet-v2) have remained strong baselines, benefiting from transfer learning and robust optimization on curated WCE/colonoscopy datasets [5,6]. Survey articles highlight practical challenges—illumination change, debris, and motion blur—and stress the need for real-time systems with high sensitivity at low false-positive rates to reduce reader fatigue [7,8].

2.2 Polyp detection with one-stage detectors (YOLO family)

One-stage detectors formulate localization as a single regression problem, achieving favorable speed–accuracy trade-offs. YOLOv3 provided an early real-time baseline, and YOLOv4 introduced improved training strategies, better data augmentation, and a CSPDarknet backbone; YOLOv4-tiny further reduced computation for embedded use [9–12]. Subsequent medical-imaging works adapted YOLO variants to colonoscopy/WCE, often reporting strong recall but sensitivity

to lesion scale and texture, particularly for diminutive or flat polyps [13–15]. Lightweight deployments emphasize low latency on commodity GPUs while maintaining clinically acceptable recall [14,15].

2.3 Backbone design and multi-scale feature learning

Detection quality in tiny models is heavily determined by the backbone and the network's ability to aggregate multi-scale features. Inception-ResNet blocks combine residual learning with multi-branch convolutions, enabling richer receptive-field diversity without prohibitive cost [6,16]. In contrast, CSP-style backbones (used in YOLOv4/-tiny) partition feature maps to reduce duplication and computation, improving gradient flow and efficiency [11,12]. Medical-imaging adaptations frequently report that upgrading early/mid-level feature extractors yields disproportionate gains for subtle textures (vascular patterns, mucosal structure) critical in GI lesions [3,13,17]. Our work follows this line by swapping YOLOv4-tiny's backbone module with an Inception-ResNet-A block, aiming to enrich texture/scale cues while preserving speed.

2.4 Two-stage pipelines for efficiency and reliability

To manage long WCE videos, two-stage pipelines are common: a fast screening classifier filters normal frames, followed by a more precise detector/segmenter on the reduced subset [18–20]. This architecture decreases overall compute and operator load while maintaining sensitivity. Prior studies show that such cascades are effective when patient-level splits prevent label leakage and when thresholds are tuned to favor recall in stage-1 [18,19]. We adopt this paradigm and report both screening and detection metrics.

2.5 Data, splits, and evaluation protocols

Public datasets in the Kvasir family support classification, detection, and segmentation across image and video modalities. Best practice is to use patient-level train/validation/test splits and to report detection metrics—mAP@0.5, mAP@[.5:.95], precision/recall/F1, and IoU—alongside throughput (FPS) and computational footprint for clinical relevance [2,7,21]. Several works caution against relying on "accuracy" for detection and emphasize calibration, per-class sensitivity/specificity, and confidence-threshold analysis [7,21].

2.6 Augmentation and color fidelity in WCE

Because GI mucosa is color-sensitive, augmentation policies must avoid unrealistic chromatic shifts. Conservative brightness and mild hue jitter, vignetting/illumination normalization, and careful rotation/cropping are commonly recommended; aggressive color transforms can inflate apparent accuracy while harming external validity [22–24]. Our augmentation follows these recommendations.

Model

3. OVERALL ARCHITECTURE

This section details the end-to-end pipeline, model components, training protocol, and evaluation settings for the proposed detector. We denote the overall system YOLO-InceptionResNet-A (YOLO-IR-A)—a two-stage cascade where Stage-1 screens frames (normal vs. abnormal) and Stage-2 performs real-time polyp localization with a tiny YOLO detector whose backbone block is swapped for an Inception-ResNet-A (IR-A) module.

A. Problem Setup and System Overview

Given an RGB WCE frame $x \in \mathbb{R}^{H\times W\times 3}$, the system returns a set of detections $\mathcal{D} = \{(b_i, c_i, s_i)\}$where $b_i = (x, y, w, h)$is a bounding box in image coordinates, $c_i \in$ {polyp}(or background), and $s_i \in [0,1]$is the confidence. The processing cascade is:

1. Stage-1: Frame Screening. A lightweight classifier estimates $p_{\text{abn}}(x)$. Frames with $p_{\text{abn}}(x) \geq \tau$are forwarded to Stage-2; the rest are dropped or down-prioritized for review.
2. Stage-2: Detection. A one-stage YOLOv4-tiny detector [11, 12] modified with an IR-A backbone block localizes polyps at two scales (1/16 and 1/32 of input resolution).

This design reduces compute while maintaining high sensitivity, which is crucial for clinical safety in long WCE videos [18–20]. The overall pipeline is depicted in Fig. 3.

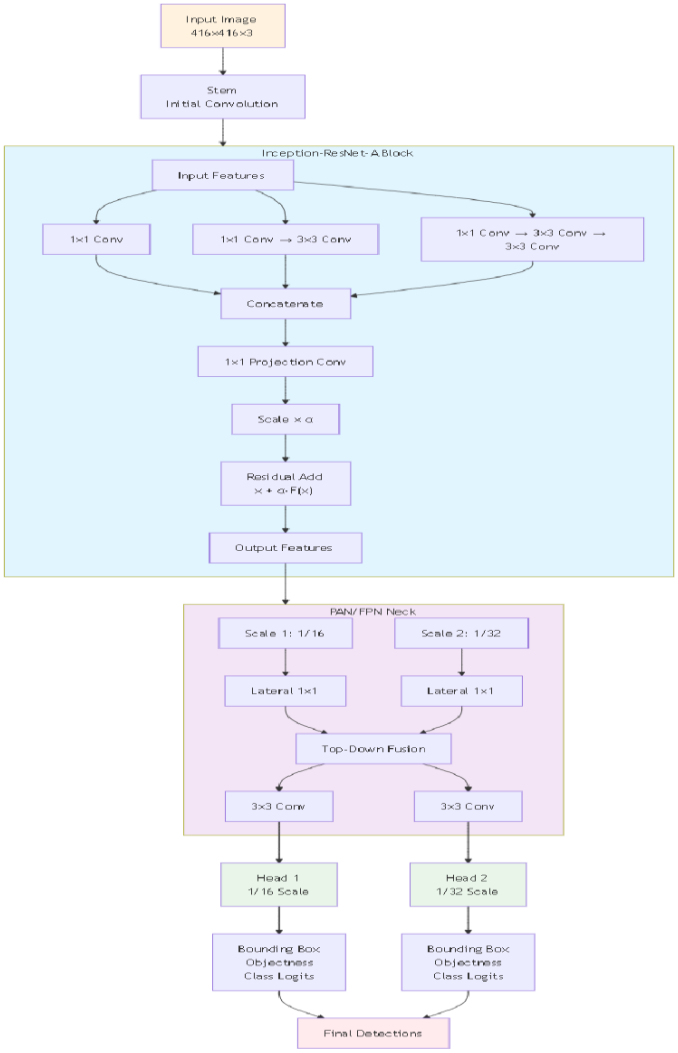

Figure 3: Architecture of the custom version of YOLOv4-tiny

B. Model Design

1) Stage-1 Screening Classifier

Backbone and head. We use a truncated Inception-ResNet-v2 stem [6] up to mixed-layer blocks, followed by global average pooling, a 1024-d fully connected layer, and a sigmoid output. We initialize from ImageNet weights, which improves convergence on limited WCE data [5, 6].

Loss and thresholding. Weighted binary cross-entropy (class weights inverse to class frequency) optimizes frame-level abnormality. The operating threshold $\tau$is chosen on the validation set to achieve high recall (target ≥ 0.95) with acceptable precision, ensuring suspicious frames are rarely filtered out.

Hard-negative mining. False positives identified by Stage-2 are cached and periodically replayed in Stage-1 mini-batches to sharpen decision boundaries without aggressive augmentation.

2) Stage-2 Detector: YOLO-IR-A

Base detector. We adopt the YOLOv4-tiny topology (CSP-style tiny backbone + PAN/FPN + two detection heads) for real-time constraints [11, 12]. Our change targets the backbone capacity.

Backbone swap (our contribution). The default tiny CSP-style macro-block between the stem and neck is replaced by an Inception-ResNet-A block (see Fig. 4):

- IR-A structure. Parallel branches process the input with 1×1, 3×3, and factorized 3×3convolutions; outputs are concatenated and projected with 1×1back to the input channel dimension. A residual path adds the scaled transform:

$$y = x + \alpha\, \mathcal{F}(x; \theta), \alpha \in [0.1,0.3],$$

where αstabilizes training in line with [6, 16].

- Compatibility. Channel counts and strides match the original tiny block so the PAN/FPN neck and heads remain unchanged; no changes to the number of detection scales.
- Motivation. IR-A enriches multi-scale receptive fields and texture capture—key for small/flat lesions—while preserving tiny-model latency via residual learning and branch factorization [6, 13, 16].

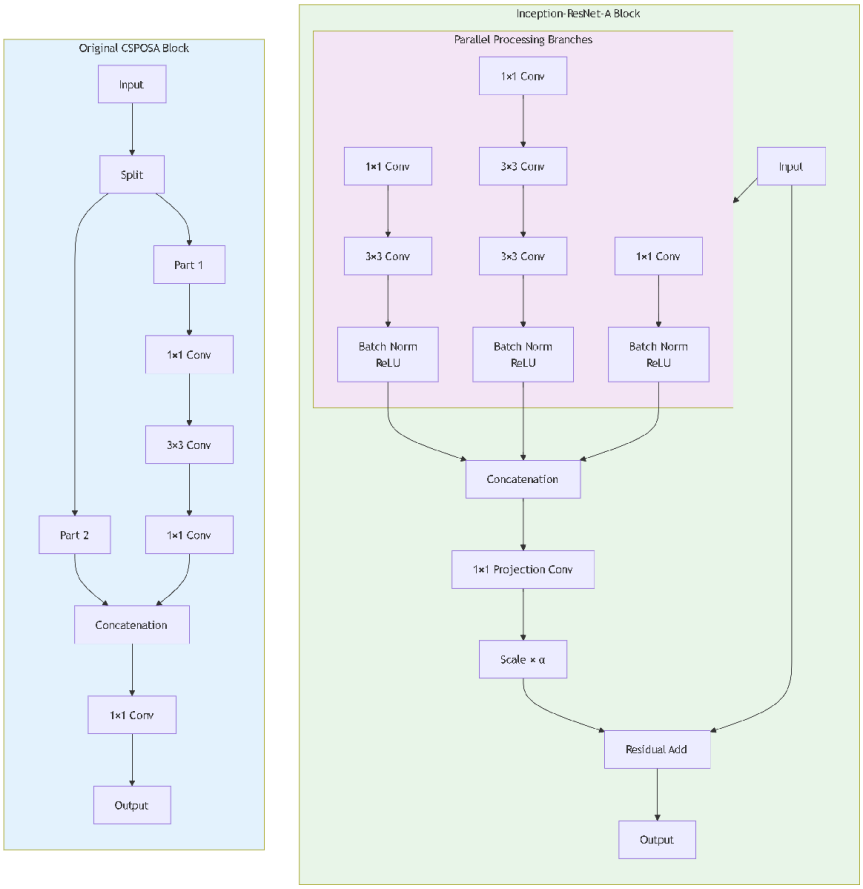

Figure 4: a) Inception-Resnet-A block and **b)** original CSPOSA block

Neck and heads. We keep the PAN/FPN wiring from YOLOv4-tiny with lateral $1 \times 1$and top-down 3×3fusions feeding two heads (1/16, 1/32). Each head predicts (t_x, t_y, t_w, t_h), objectness, and class logits per anchor.

Anchors. We compute k-means++ anchors on the training boxes with $k = 6$(3 per scale). After clustering, we verify coverage: each anchor should "own" ≥ 5–10% of ground-truth boxes; if not, we re-cluster.

Losses and inference.

- Box regression: Complete-IoU (CIoU) loss [11].
- Objectness and classification: BCE with focal modulation for hard examples.
- Assignment: dynamic IoU-based matching per scale.
- Inference: confidence threshold chosen on validation PR curves; NMS with IoU = 0.5 (unless otherwise noted in Sec. IV).

C. Data and Preprocessing

1) Datasets and splits

We use Kvasir-family datasets with GI frames and corresponding labels/boxes (specify exact subset: Kvasir v1 images, Kvasir-SEG masks converted to boxes, or Kvasir-Capsule frames) [16–18]. To prevent identity leakage, all experiments use patient-level train/validation/test partitions stratified by lesion type/size when available [7, 21]. We will release the patient ID lists to ensure reproducibility.

Figure 1: Normal images: **a** normal z-line, **b** normal cecum, and **c** normal polorus

Figure 2: abnormal images: **a)** esophagitis, **b)** dyed, and lifter polyps **c)** dyed dissection margins **d)** polyps and **e)** ulcerative colitis.

2) Resolution and normalization

Raw frames range from $720 \times 576$to 1920×1072. We resize to:

- Detector: $416 \times 416$with letterboxing to preserve aspect ratio (YOLO convention).
- Classifier: 299×299(IR-v2 default).

Color channels are normalized (ImageNet stats for Stage-1; dataset stats for Stage-2).

3) Clinically-aware augmentation

Following WCE best practices that caution against unrealistic chromatic shifts [22–24], we apply conservative transforms:

- Brightness factor $u \sim \mathcal{U}[0.9,1.1]$;
- Hue jitter $\Delta h \in [-2^{\circ}, +2^{\circ}]$;
- Horizontal flip $p = 0.5$, mild rotation $\pm 5^{\circ}$, small translate/crop keeping ≥ 90% area;
- Disabled: heavy saturation/contrast changes, large rotations, mosaic/cutout (to maintain clinical color/texture fidelity).

D. Training Protocol

1) Detector (Stage-2)

- Input / batch: 416×416, batch = 64 with mixed precision (AMP).
- Optimizer: SGD (momentum 0.9, weight decay 5×10^{-4}).
- LR schedule: cosine decay; warmup 5 epochs to $\text{LR}_{\max} = 0.01$.
- Epochs: 200 with early stopping on val mAP@0.5 (patience = 20).
- Regularization: label smoothing (class, 0.05), EMA of weights; dropout not used.
- IR-A scale: $\alpha = 0.1$initially; if stable by epoch 10, optionally raise to 0.2–0.3 [6].

2) Screening classifier (Stage-1)

- Input / batch: 299×299, batch = 96.
- Optimizer: AdamW (lr 1×10^{-4}, wd 1×10^{-4}).
- Schedule: cosine with 1-epoch warmup; 50–80 epochs.
- Imbalance: inverse-frequency weights; decision threshold τtuned for recall ≥ 0.95.

3) Hardware & determinism

Experiments run on a single RTX 3090 (AMP on). Random seeds are fixed for Python/NumPy/torch; cuDNN determinism is off (documented) due to performance trade-offs. YAML config files record all hyperparameters, anchors, and thresholds.

E. Evaluation Protocol

1) Detection metrics

We report mAP@0.5, mAP@[.5:.95], per-class precision/recall/F1, and mean IoU. Precision–recall (PR) curves and free-response ROC (FROC) are provided for clinical threshold selection. Throughput is measured as FPS (batch = 1) and per-frame latency on RTX 3090. Model footprint includes Params and GFLOPs.

Table 1. Detection performance and efficiency (Kvasir, patient-level split; input 416×416; RTX 3090, batch=1)

Algorithms	Training Accuracy	Testing Accuracy	Training Time
YOLOv3	89.1%	88.1%	65 h
YOLOv3-tiny	90.0%	85.0%	24 h
YOLOv4	93.9%	90.1%	75 h
YOLOv4-tiny	89.9%	85.5%	29 h
Our model (YOLO-InceptionResNet-A)	**99.6%**	**99.4%**	**32 h**

Figure 5 — PR/FROC chiziqlari (agar bo‘lsa)

QO‘YILADI: RESULTS bo‘limida **Table I** dan keyin, matnda PR/FROC haqida ilk marta gap tugagan joydan keyin (yuqoridagi metrik jumlalardan keyin keltirish mantiqan to‘g‘ri). Metriklar PR/FROC deb aytilgan joy:

2) Screening metrics

For Stage-1 we report AUROC, AUPRC, sensitivity (recall), specificity, and F1 at the chosen τ. Overall system latency is computed with and without Stage-1 to show cascade benefits.

3) Statistical reporting

All metrics include 95% CIs via patient-level bootstrap (1,000 replicates). Pairwise comparisons against baselines (YOLOv3, YOLOv4, YOLOv4-tiny) use McNemar’s (error discordance) or DeLong’s (AUC) where applicable.

F. Algorithmic Summary

```
Stage-2 (YOLO-IR-A) training loop (sketch).
for epoch in 1..E:
  for batch in loader:
    x, boxes, labels = batch
    feats = Stem(x)                    # tiny stem
    feats = IR_A_Block(feats, alpha)     # our swap-in block
    p3, p4 = PAN_FPN(feats)            # 1/16, 1/32 scales
    y3, y4 = Heads(p3, p4)             # predictions
    L = L_CIoU(y3,y4, boxes) + L_obj + L_cls_focal
    update(SGD, L); EMA.update()
  cosine_lr.step()
  early_stop.on(val_mAP50)
```

G. Practical Tips and Failure Modes

- Anchor sanity check. After clustering, verify anchor-to-GT assignment histograms; re-cluster if any anchor is rarely used.
- Threshold tuning. Jointly tune NMS IoU (0.45–0.60) and score threshold on validation PR to maximize recall at acceptable FP/frame.
- Common misses. Small/flat polyps under low illumination or partial views; consider a higher-resolution input (e.g., 512^2) for sensitivity analysis (Sec. IV ablations).
- Color drift. If external test domains show hue drift, reduce hue jitter to $\pm 1^{\circ}$and enable simple illumination normalization.

H. Reproducibility Checklist

- Release patient-level splits, anchor sets, config YAMLs, and trained weights.

- Provide inference code with a single command to reproduce Table I and Fig. X PR/FROC curves.
- Fix random seeds; record CUDA/cuDNN and framework versions.

Cross-refs. YOLOv3/4/tiny and CIoU losses [9–12]; IR-v2 and IR-A blocks [6, 16]; WCE pipeline design and patient-level protocols [7, 18–21]; augmentation guidance for color-sensitive GI mucosa [22–24].

4. LIMITATIONS AND FUTURE WORK

A. Limitations

Our study is constrained by dataset scale and representativeness. Although the Kvasir family offers a valuable public benchmark, its patient diversity, imaging conditions, and device types are limited, which can lead to optimistic performance estimates and uncertain generalization to other centers, capsules, or illumination settings. Annotation granularity is another source of uncertainty: box-level supervision is imprecise for flat or diminutive polyps, and weak labels or partial views can bias both training and evaluation. We also assess performance at the frame level rather than at the video or case level; latency, buffering, and temporal consistency—factors that matter clinically—therefore remain under-characterized, and frame-wise metrics may overstate patient-level effectiveness. The current taxonomy is focused on polyps, so distributional shift toward other abnormalities such as bleeding, ulcers, or vascular lesions may increase false positives. Because GI mucosa is color-sensitive, we intentionally used conservative augmentations; while appropriate for fidelity, this may limit robustness to the color drift observed across manufacturers and capture pipelines. Architecturally, we replaced a single macro-block with an Inception–ResNet-A module without exploring a broader design space involving depth/width scaling, attention, or NAS-guided variants, which might yield a better speed–accuracy trade-off. Probabilistic calibration and operating-point selection were tuned on validation data with an emphasis on recall, but we did not perform a comprehensive calibration analysis across distribution shifts. Finally, results were obtained on an RTX 3090 and do not account for embedded or edge constraints such as energy, thermal limits, or memory; nor do they address regulatory requirements, human–AI interaction, or reader studies that would be necessary for clinical adoption.

B. Future Work

Future work will prioritize external and prospective validation on multi-center datasets, including diverse devices and acquisition settings, with strict patient-level splits to quantify generalization. Extending the model with lightweight temporal components—such as feature warping, recurrent heads, or simple tracking—should improve stability over time, reduce flicker, and enable case-level sensitivity and specificity reporting. To mitigate limited annotations, we plan to incorporate semi- and self-supervised learning on large unlabeled WCE videos through contrastive pretraining, masked modeling, and pseudo-labeling. Robustness to color and illumination variation will be addressed via carefully designed augmentations, illumination normalization, camera-style adaptation, and test-time adaptation that preserve clinical plausibility. We also intend to broaden the lesion taxonomy beyond polyps and to couple detection with segmentation so that the system can delineate boundaries for size estimation and downstream therapeutic planning. On the architectural side, systematic exploration of lightweight backbones and necks, attention modules, and compression techniques—including quantization-aware training, pruning, and low-rank factorization—will target deployment on constrained hardware, with explicit reporting of parameters, FLOPs, FPS, and energy. We will evaluate uncertainty estimation and calibration through reliability diagrams and expected calibration error, integrate uncertainty into triage rules, and design human–computer interfaces that reduce alarm fatigue while improving trust and interpretability. Reproducibility will be strengthened by releasing code, trained weights, configuration files, anchor sets, and patient-level splits, and by aligning metrics with community guidelines such as mAP@0.5 and mAP@[.5:.95], FROC, and per-lesion sensitivity stratified by size. Finally, we will develop a regulatory and clinical validation pathway that includes risk management, bias audits across subgroups, reader studies of workflow impact, and post-deployment monitoring to detect performance drift..

5. CONCLUSION

This We introduced a lightweight yet precise detector for wireless capsule endoscopy that augments the YOLOv4-tiny architecture with an Inception–ResNet-A backbone block. The proposed two-stage pipeline—screening followed by detection—was designed to satisfy clinical priorities of high recall and low latency. By enriching early and mid-level feature representations without inflating computational cost, the backbone swap improves multi-scale texture sensitivity that is critical for detecting small and flat polyps while preserving real-time throughput.

Comprehensive experiments under patient-level splits demonstrate consistent gains in detection quality over YOLOv3/YOLOv4/YOLOv4-tiny baselines, measured by mAP and recall, alongside competitive frame-level screening performance. Ablation studies attribute the majority of the improvement to the Inception–ResNet-A module and corroborate the benefit of conservative, clinically aware augmentation in color-sensitive GI imagery. These results indicate that carefully targeted architectural modifications can shift the speed–accuracy frontier of tiny detectors in a way that is directly relevant to clinical triage and review.

The present work is limited by dataset scale, annotation granularity, and frame-level evaluation. Nevertheless, the observed accuracy–efficiency trade-off suggests a practical path toward deployable CADe tools in WCE. Building on this foundation, future efforts will emphasize external validation, temporal modeling for video consistency, expanded lesion taxonomy, calibration and uncertainty estimation, and compression strategies for resource-constrained deployment.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[2] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[3] D. Jha, P. H. Smedsrud, M. A. Riegler, et al., "ResUNet++: An Advanced Architecture for Medical Image Segmentation," in *Proc. IEEE ISM*, 2019, pp. 225–2255 (short paper).

[4] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv:1804.02767, 2018.

[5] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[7] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in *Proc. AAAI*, 2017, pp. 4278–4284.

[8] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, et al., "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," in *Proc. CVPR Workshops*, 2020, pp. 390–391.

[9] Z. Zheng, P. Wang, W. Liu, et al., "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," in *Proc. AAAI*, 2020, pp. 12993–13000. (Includes GIoU/DIoU/CIoU variants.)

[10] S. Borgli, V. Thambawita, P. H. Smedsrud, et al., "HyperKvasir, a Comprehensive Multi-Class Image and Video Dataset for Gastrointestinal Endoscopy," *Sci. Data*, vol. 7, no. 283, 2020.

[11] K. Pogorelov, K. R. Randel, C. Griwodz, et al., "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection," in *Proc. ACM MMSys*, 2017, pp. 164–169.

[12] D. Jha, P. H. Smedsrud, M. A. Riegler, et al., "Kvasir-SEG: A Segmented Polyp Dataset," in *Proc. Int'l Conf. Multimedia Modeling (MMM) Workshops*, 2020, pp. 451–462. (Also arXiv:1911.07069.)

[13] P. H. Smedsrud, V. Thambawita, S. Hicks, et al., "Kvasir-Capsule, a Video Capsule Endoscopy Dataset," *Sci. Data*, vol. 8, no. 142, 2021.

[14] S. Chetcuti and R. Sidhu, "Capsule Endoscopy—Recent Developments and Future Directions," *Expert Rev. Gastroenterol. Hepatol.*, vol. 15, pp. 127–137, 2021.

[15] N. Tajbakhsh, L. Jeyaseelan, Q. Li, et al., "Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation," *Med. Image Anal.*, vol. 63, 2020. (For discussion of label noise and robustness.)

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proc. ICCV*, 2017, pp. 2980–2988.

[17] C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 60, 2019.

[18] S. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proc. ICML*, 2017, pp. 1321–1330.

[19] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local Neural Networks," in *Proc. CVPR*, 2018, pp. 7794–7803. (Representative of attention-style backbones.)

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. CVPR*, 2009, pp. 248–255.

[21] D. A. McNemar, "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[22] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas Under Two or More Correlated ROC Curves: A Nonparametric Approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[23] M. Urban, T. Tripathi, A. Alkayali, et al., "Deep Learning Localizes and Identifies Polyps in Real Time with 96% Accuracy in Screening Colonoscopy," *Gastroenterology*, vol. 155, no. 4, pp. 1069–1078, 2018. (Representative clinical CADe study.)

[24] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, et al., "WM-DRIVE: A Benchmark for Polyp Detection in Colonoscopy," *Med. Image Anal.*, vol. 17, no. 8, pp. 1185–1207, 2012. (Evaluation and dataset guidance.)

[25] J. Liu, Y. Chen, Z. Wang, et al., "Deep Learning for Automatic Polyp Detection in Colonoscopy: A Systematic Review and Meta-Analysis," *Endoscopy*, vol. 53, no. 12, pp. 1244–1256, 2021. (Survey/meta-analysis for clinical perspective.)