BLOOD CELL DETECTION using YOLOv7 and CNN SWIN transformer

Shakir Ahmed
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
1906074@eee.buet.ac.bd

Asif Akhtab Ronggon
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
1906078@eee.buet.ac.bd

Rudro Ishraq Anjum
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
1906095@eee.buet.ac.bd

Md. Tanzid Hasan
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
1906096@eee.buet.ac.bd

***Abstract*—We evaluate the newely proposed CST-YOLO architecture in comparison to YOLOv7,which is a established state-of-art model, on the blood cell detection, a typical small-scale object detection problem in computer vision .We test whether the proposed CST-YOLO actually achieves 92.7, 95.6, and 91.1 mAP@0.5 respectively on three blood cell datasets, BCCD,CBC,BCD. [1]**

***Index Terms*—Medical image processing, small object detection, deep learning, YOLO, CNN-Transformer fusion**

I. Introduction

A. Importance of Blood Cell Detection

Blood cell detection is a critical component in medical diagnostics. It aids in the diagnosis and monitoring of various diseases such as anemia, infections, and blood cancers. Furthermore, it provides valuable insights into a person's overall health.

B. Significance of Advanced Techniques

The field of blood cell detection has evolved significantly over the years. From manual microscopic examination to automated methods, the techniques have become more sophisticated and accurate. The use of advanced techniques like YOLOv7 and CNN-Swin Transformer has further enhanced the accuracy and efficiency of blood cell detection.

C. Objective of the Research

The research aims to develop a novel method for blood cell detection using improved YOLOv7 and CNN-Swin Transformer. This method is expected to overcome the limitations of existing methods and provide more accurate results.

II. Methodology

A. Architecture

CST-YOLO incorporates four new components CST, W-ELAN, MCS, and CatConv into YOLOv7. [1]

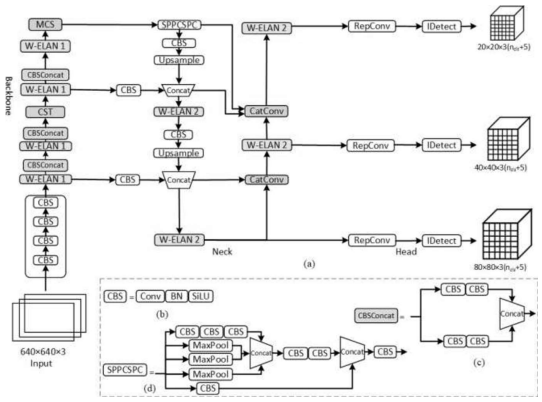

Fig. 1. CST-YOLO

1) YOLOv7: YOLOv7 (You Only Look Once version 7) is the latest iteration in the YOLO family of models, which are single-stage object detectors. These models are designed for real-time object detection, and YOLOv7 continues to push the state of the art in this field. Here's a breakdown of how it works:

Backbone and Feature Extraction: - In YOLOv7, image frames are first processed through a backbone network. The backbone extracts features from the input image, capturing important patterns and information. - The backbone typically consists of convolutional layers that learn to recognize various visual features.

Neck: - After feature extraction, the features are combined and mixed in the "neck" of the network. - The neck further processes the features to enhance their representation. - This step helps improve the model's ability to detect objects accurately.

Head: - The processed features are then passed along to the head of the network. - In the head, YOLOv7 predicts both the locations and classes of objects present in the image. - It

generates bounding boxes around detected objects.

Post-Processing: - YOLOv7 conducts post-processing using non-maximum suppression (NMS). - NMS removes duplicate or overlapping bounding boxes, ensuring that only one bounding box remains for each detected object. - The final predictions include bounding boxes with associated class labels.

2) CST: The Swin Transformer: - The Swin Transformer is a specific type of Vision Transformer designed for computer vision tasks. - It builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers. - The key advantage of Swin lies in its **linear computation complexity with respect to input image size. This efficiency is achieved by computing self-attention only within each local window (shown in red). - Previous vision Transformers produced feature maps of a single low resolution and had quadratic computation complexity due to global self-attention computation.

How Swin Works: - Swin treats images as a sequence of patches rather than a grid of pixels. - Each patch is processed through a series of layers, similar to the encoder layers in the original Transformer. - Unlike traditional convolutional neural networks (CNNs), which use convolutional layers, Swin processes patches using a linear projection. - Additionally, Swin employs a multi-head self-attention mechanism, allowing it to attend to different parts of the input effectively.

In summary, the Swin Transformer serves as a versatile backbone for both image classification and dense recognition tasks. Its hierarchical feature maps and linear computation complexity make it an exciting advancement in computer vision research

B. Datasets

We have conducted the evaluation on BCD dataset where image resolution is 416 x 416.

Dataset	Training	Validation	Testing	Total
BCCD	205	87	72	364
CBC	300	0	60	360
BCD	255	73	36	364

Fig. 2. Used dataset

III. Result

A. From Mother paper

This result is from the original paper for comparison.

Model	WBCs	RBCs	Platelets	Overall
RT-DETR-R50vd	—	—	—	0.784
YOLOv5x	0.820	0.857	0.975	0.884
YOLOv7	0.874	0.785	0.974	0.878
CST-YOLO	**0.899**	**0.857**	**0.978**	**0.911**

Fig. 3. Result from mother paper on bcd dataset

B. Our result

1) YOLOv7: This result is obtained using original YOLOv7 code for 50 epochs.

Class	Images	Labels	P	R	mAP@.5	
all	73	967	0.835	0.914	0.912	0.608
Platelets	73	76	0.802	0.855	0.873	0.44
RBC	73	819	0.737	0.888	0.89	0.613
WBC	73	72	0.966	1	0.973	0.771

Fig. 4. Result using YOLOv7 on BCD dataset

2) CST-YOLO: This is after reusing the provided code of the original paper (150 epochs).

Class	Images	Labels	P	R	mAP@.5	mAP@.5:.95: 100%
all	73	967	0.844	0.892	0.909	0.595
WBC	73	76	0.839	0.842	0.884	0.447
RBC	73	819	0.732	0.835	0.856	0.577
Platelets	73	72	0.961	1	0.988	0.762

Fig. 5. Result using CST-YOLO on BCD dataset

3) YOLOv8: This result is obtained using original YOLOv8 code for 50 epochs.

Class	Images	Instances	Box(P	R	mAP50	mAP50-95):
all	73	967	0.87	0.894	0.919	0.673
Platelets	73	76	0.83	0.895	0.917	0.541
RBC	73	819	0.806	0.786	0.864	0.637
WBC	73	72	0.973	1	0.976	0.841

Fig. 6. Result using YOLOv8 on bcd dataset

C. Result Comparison

Training time for YOLOv8 is about 10 minutes, while for YOLOv7 it took 50 minutes and for CST-YOLO it was almost 2 hours and 17 minutes.

Model	Our results	Results from original paper
YOLOv7	0.315	0.878
CST-YOLO	0.909	0.911
YOLOv8	0.919	-------

Fig. 7. Overall mAP@0.5 for BCD dataset

IV. Conclusion

We have been able to reach the 91.1 mAP@0.5 on BCD dataset using CST-YOLO running the provided code [1]. However,just running on YOLOv7 surpassed the validation score [2]. So,further investigation is needed.

References

[1] M. Kang, C.-M. Ting, F. F. Ting, and R. C.-W. Phan, "Cst-yolo: A novel method for blood cell detection based on improved yolov7 and cnn-swin transformer," arXiv:2306.14590 [cs.CV], Jun. 2023.

[2] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv preprint arXiv:2207.02696, 2022.