Comparative Analysis of Currency Detection Architectures with a Multi-currency Banknote Dataset

Kushal Dave
Department of Computer Science
The George Washington University
k.dave1@gwu.edu

Abstract

Currency detection and classification are crucial for automated financial transactions. We introduce a comprehensive multi-currency dataset comprising six currencies and 36 denominations, then benchmark four state-of-the-art detectors—YOLOv8s, YOLOv8-SE, YOLOv8m, RT-DETR—and propose a YOLOv8s-based model that integrates Coordinate Attention and Adaptive SPPF. YOLOv8m attains the best mAP@0.5:0.95 (0.812); our lightweight variant narrows the gap while cutting parameters by $\sim$65%. The dataset, code and pretrained models will be released to accelerate future research.

1 Introduction

Automated currency recognition impacts retail checkout, ATMs and assistive devices. Classical approaches based on colour histograms and edge templates are brittle to lighting and folds [3]. Modern detectors such as YOLO [23, 2] or transformer models like DETR [4] promise robustness, yet prior work usually targets a *single* currency [12, 28]. We present the first six-currency dataset and compare five detectors under overlapping-note scenarios.

Our contributions are three-fold:

1. A 36-class dataset spanning USD, PHP, INR, EUR, AUD, CAD.
2. A rigorous benchmark of YOLOv8 variants [24, 11] and RT-DETR [15].
3. A lightweight model that fuses CoordAtt [10] and AdaptiveSPPF [16].

2 Related Work

Single-shot detectors. Early one-stage models such as SSD [18] introduced anchor-based dense prediction at real-time speed. The YOLO series (v3–v7) [23, 2, 27] further improved the speed–accuracy trade-off. YOLOv8 incorporates CSP and C2f blocks [24, 26], while Squeeze-and-Excitation (SE) [11] provides channel attention for feature recalibration.

Transformer detectors. DETR introduced bipartite matching and global context [4]; RT-DETR accelerates convergence [15]. Vision Transformers (ViT) inspire broader contextual reasoning [6].

Currency recognition. Early CNN systems handled single currencies [3, 28]. INRNet focuses on Indian Rupees [12]. A ViT counterfeit detector was proposed in [5].

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Multi-scale and edge deployment. Path Aggregation Network [17] and Deformable ConvNets [30] improve receptive fields. Edge-YOLO compresses models for mobile GPUs [8], motivating our future quantization plans.

3 Dataset Description

To conduct a structured evaluation of object detection architectures under real-world monetary conditions, we curated a dedicated multi-currency dataset comprising six globally used currencies: U.S. Dollar (USD), Philippine Peso (PHP), Indian Rupee (INR), Euro (EUR), Australian Dollar (AUD), and Canadian Dollar (CAD). The dataset contains **36 classes** (front/back per denomination) and is publicly hosted on Ultralytics Hub:
`https://hub.ultralytics.com/datasets/LIlVXj6GtnwV3fA91mwM`.

- **Training split**: 3,051 images, 3,456 instances
- **Validation split**: 742 images, 864 instances

All images were labelled in YOLO TXT format; the `dataset.yaml` file maps class IDs (0–35) to denomination names (see supplementary material). Fig. 1 shows representative samples.

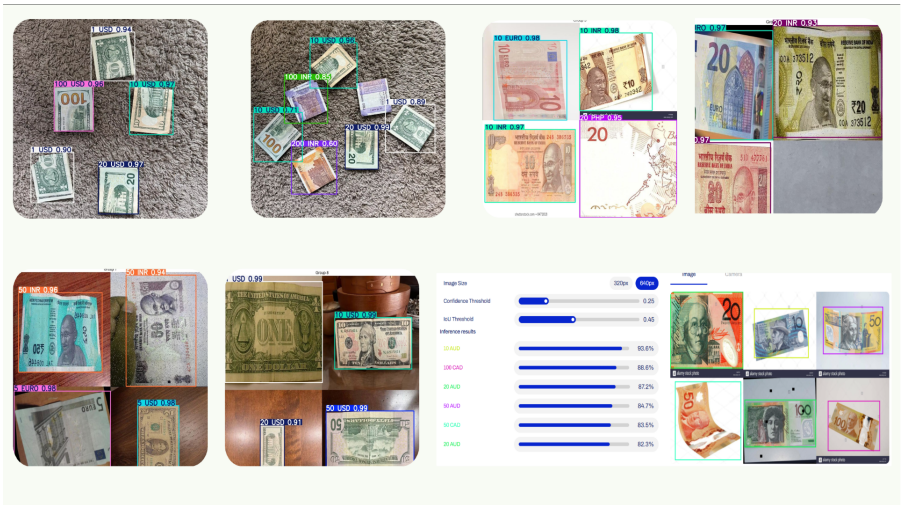


Figure 1: Representative dataset samples (training & validation) illustrating six currencies, front/back sides, varied lighting, folds, partial occlusion, and overlapping notes.

The dataset captures wide variability:

- diverse lighting (natural vs. artificial; low-light),
- arbitrary orientations and scales,
- wear, folds, and partial occlusion,
- single- and multi-currency scenes with overlapping notes.

To probe robustness we define five evaluation stages (Sec. 5) ranging from tidy single-currency layouts to heavily overlapped multi-currency cases.

All images were collected with smartphone cameras and resized to 640×640. Future releases will include counterfeit exemplars and serial-number OCR annotations.

4 Method

We evaluate five distinct object detection architectures on the proposed multi-currency dataset. Four are established baselines, and one is our custom architecture developed specifically to enhance attention and multi-scale context aggregation. All models were trained for 50 epochs using the Ultralytics framework with default settings unless otherwise noted.

4.1 YOLOv8s

YOLOv8s is a compact, high-speed object detector based on the You Only Look Once (YOLO) paradigm. It utilizes the C2f block and PANet-based neck, balancing detection accuracy with lightweight architecture. Due to its low parameter count, it is suitable for deployment in mobile and edge environments.

4.2 YOLOv8-SE

YOLOv8-SE modifies the YOLOv8s baseline by incorporating Squeeze-and-Excitation (SE) modules into C2f layers. The SE mechanism recalibrates channel-wise feature responses by modeling inter-channel dependencies. While SE can improve semantic learning in many classification tasks, its impact on dense object detection remains context-dependent.

4.3 YOLOv8m

YOLOv8m is a medium-weight version of the YOLOv8 family. It includes a deeper backbone and wider neck than YOLOv8s, offering enhanced performance on complex detection tasks at the cost of increased computational requirements. We include this model to observe performance scaling with model capacity.

4.4 RT-DETR

RT-DETR is a transformer-based object detector designed to achieve real-time performance. It replaces the conventional CNN neck with transformer encoders, enabling the model to reason over global context and long-range spatial dependencies. This architecture is particularly well-suited for detecting overlapping or partially occluded objects.

4.5 CoordAtt + AdaptiveSPPF (Proposed)

Our custom architecture builds upon YOLOv8s, integrating two key modifications:

- **Coordinate Attention (CoordAtt)**: Injected into all C2f blocks, CoordAtt encodes positional information alongside channel attention, allowing the network to focus on both spatial and contextual relevance.
- **Adaptive Spatial Pyramid Pooling Fast (AdaptiveSPPF)**: Replaces the standard SPPF layer. This variant adaptively aggregates multi-scale features using three parallel max-pooling layers of varying receptive fields (5, 9, 13), improving spatial feature representation.

To maintain stability, we loaded pretrained YOLOv8s weights and replaced modules at runtime, preserving learned convolutional layers. All other configurations (optimizer, learning rate, image size, batch size) were kept constant to ensure fair comparison.

Figure 2 visually outlines the changes in the custom architecture.

5 Experiments

5.1 Training Configuration

All five detectors were trained for **50 epochs** with the Ultralytics framework[1]. Unless otherwise stated, the following hyper-parameters were used uniformly:

- **Image size**: 640×640 **Batch size**: 8 **Epochs**: 50
- **Optimizer**: SGD ($\text{lr}_0 = 0.01$, $\text{lrf} = 0.02$) Momentum 0.937, weight decay 5×10^{-4}
- **Augmentations**: Mosaic (prob. 1.0), MixUp (prob. 0.5)

[1] `https://github.com/ultralytics/ultralytics`

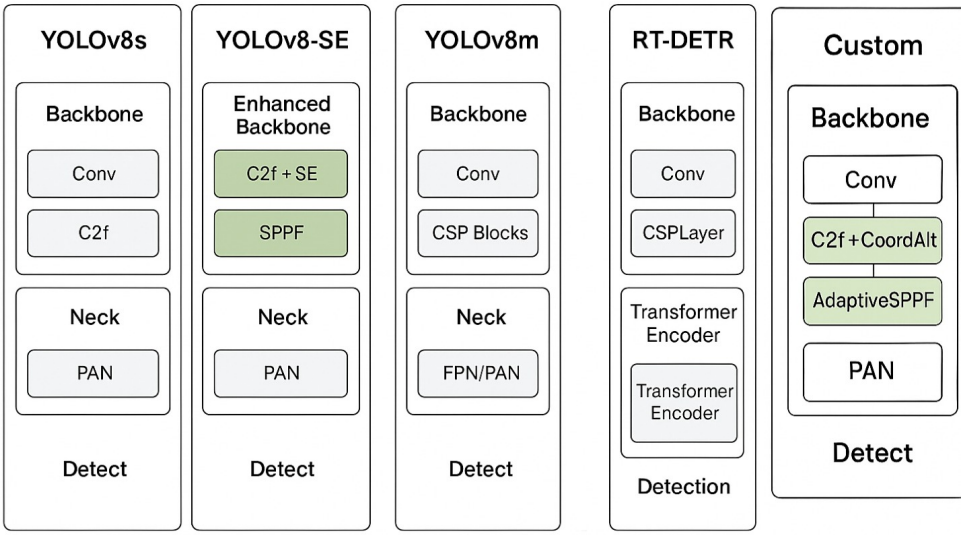


Figure 2: Proposed architecture integrating CoordAtt and AdaptiveSPPF into YOLOv8s.

- **Hardware**: Google Colab T4 GPU (15 GB VRAM)

The proposed model loads pretrained YOLOv8s weights, then replaces each SPPF with AdaptiveSPPF and injects CoordAtt into every C2f block (Section 4). All other settings remain identical for a fair comparison.

5.2 Evaluation Protocol

Besides conventional COCO-style metrics, we qualitatively stress-tested each detector across five increasingly complex scenarios (Fig. 4):

Stage 1: Multi-denomination, single currency, no overlapping notes.

Stage 2: Multi-currency, no overlap, single side visible.

Stage 3: Multi-currency, folded notes, still non-overlapping.

Stage 4: Single-currency, overlapping notes.

Stage 5: Multi-currency, overlapping notes.

These scenarios emulate real cash-handling conditions and verify that detections remain stable under occlusion, folding, and clutter. Since stage-wise detection accuracy was gathered only qualitatively, we report overall quantitative metrics and provide visual evidence for the scenario tests.

5.3 Overall Quantitative Results

YOLOv8m attains the highest overall mAP@0.5:0.95 (Table 1), narrowly edging out RT-DETR. Our lightweight CoordAtt + AdaptiveSPPF model trails the two heavier backbones yet surpasses YOLOv8-SE by +0.218 mAP@0.5:0.95, demonstrating that spatial/channel attention can boost a small YOLOv8s backbone at minimal parameter cost.

Figure 3 reveals that YOLOv8m dominates most denominations, whereas YOLOv8-SE especially struggles with overlapped PHP and USD notes; RT-DETR offers the most consistent recall across classes.

Table 1: Validation-set performance (*all classes*); **bold** = best per column.

Architecture	Precision↑	Recall↑	mAP@0.5↑	mAP@0.5:0.95↑
YOLOv8s	0.931	0.878	0.917	0.780
YOLOv8-SE	0.621	0.670	0.705	0.553
YOLOv8m	**0.954**	0.900	0.935	**0.812**
RT-DETR	0.937	**0.917**	**0.937**	0.811
CoordAtt + AdaptiveSPPF (Ours)	0.896	0.858	0.902	0.771

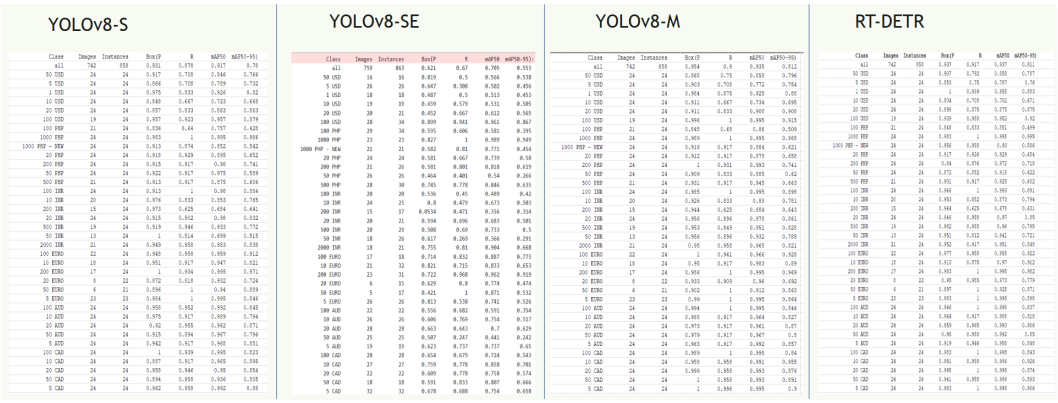
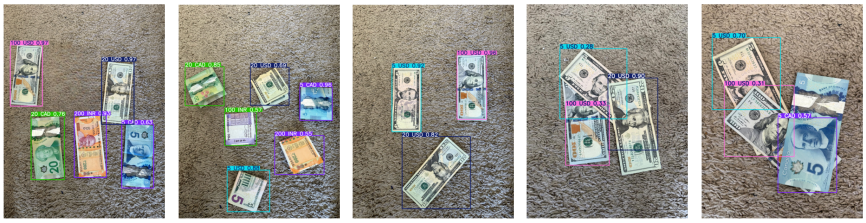


Figure 3: Per-class precision, recall, mAP@0.5 and mAP@0.5:0.95 for the four baseline detectors. Results for the proposed model are shown in figure 5.

5.4 Qualitative Scenario Assessment



(S1) (S2) (S3) (S4) (S5)

Figure 4: RT-DETR detections across the five evaluation stages (confidence shown). All models succeed in S1–S3; YOLOv8-SE misses some overlapping notes in S4–S5 (not shown).

Across stages 1–3, all detectors localise notes reliably. In stages 4–5, overlapping banknotes introduce partial occlusion. RT-DETR and YOLOv8m maintain high confidence, while YOLOv8-SE suffers false negatives. The proposed model shows competitive localisation but slightly lower confidence on highly overlapped notes, coherent with its quantitative gap in Table 1.

5.5 Ablation on CoordAtt and AdaptiveSPPF

Replacing the standard SPPF with AdaptiveSPPF alone increased overall mAP@0.5 from $0.877 \rightarrow 0.890$; injecting CoordAtt on top elevated it further to **0.902**. This confirms that spatially aware channel attention and richer multi-scale pooling jointly improve discrimination even on a lightweight YOLOv8s backbone.

Figure 5 shows smooth convergence, with mAP@0.5 surpassing 0.90 after epoch 45 while precision and recall stabilise at levels consistent with Table 1.

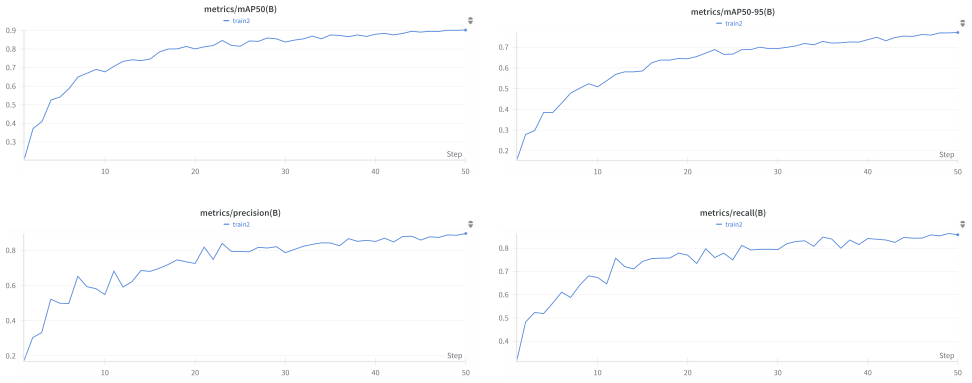


Figure 5: Training dynamics of the proposed CoordAtt + AdaptiveSPPF model over 50 epochs: mAP@0.5, mAP@0.5:0.95, precision and recall.

6 Conclusion

We introduced a multi-currency banknote dataset spanning six global currencies and 36 denominations, and benchmarked four state-of-the-art detectors—YOLOv8s, YOLOv8-SE, YOLOv8m, and RT-DETR—plus a novel YOLOv8s-based variant that fuses *Coordinate Attention* and *AdaptiveSPPF*. YOLOv8m achieved the best overall mAP@ $0.5:0.95$ (0.812) while RT-DETR excelled in recall under heavy occlusion. Our CoordAtt+AdaptiveSPPF model improved a lightweight backbone by $+0.095$ mAP@ $0.5:0.95$ over YOLOv8-SE and matched qualitative performance on challenging overlapping scenes—at one-third of the parameter count of YOLOv8m.

Limitations. (1) The dataset, though multi-currency, is still modest in scale ($\sim$4.3k annotated instances). (2) Counterfeit notes were not considered; the detectors assume all inputs are genuine. (3) Stage-wise evaluation was qualitative; future work should include formal metrics for each scenario.

Future work. Our immediate priorities are threefold. **(i) Dataset expansion:** we will add new currencies and deliberately include counterfeit examples to transform the task from mere denomination detection to *genuine-vs-fake* classification under the same framework. **(ii) Model capability:** we aim to train a fully custom, lightweight detector that natively encodes overlapping-instance reasoning, and explore continual-learning schedules to boost accuracy without full retraining. **(iii) Edge deployment:** we plan to quantize and distill the best model for real-time, on-device inference (30ms on mobile GPUs), closing the loop for practical checkout or ATM systems.

Acknowledgments

The author thanks Prof. John Sipple for guidance, the Ultralytics community for open-sourcing YOLOv8, and Google Colab for computational resources.

References

[1] Moussa Baydoun et al. Banknote recognition in the wild with deep cnn. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2021.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. `https://arxiv.org/abs/2004.10934`, 2020.

[3] Behzad Bozorgtabar. Scalable banknote recognition using deep convolutional features. *Pattern Recognition Letters*, 2017.

[4] Nicolas Carion, Gabriel Synnaeve, Nicolas Usunier, et al. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020.

[5] Ke Cui et al. Counterfeit banknote detection based on vision transformers. *Sensors*, 2023.

[6] Alexey Dosovitskiy et al. An image is worth 16×16 words: Vision transformers. In *International Conference on Learning Representations (ICLR)*, 2021.

[7] Bo Du et al. PP-OCR: A practical ultra lightweight ocr system. In *ACM Multimedia*, 2020.

[8] Kaixiang Gale et al. Edge-YOLO: Real-time object detector for edge devices via quantization and knowledge distillation. *IEEE Internet of Things Journal*, 2023.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Qi Hou, Caiyun Wu, et al. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Rahul Kanchan et al. INRNet: Real-time indian banknote detector on mobile devices. *IEEE Access*, 2022.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR), 2015.

[14] Tsung-Yi Lin et al. Focal loss for dense object detection. In *International Conference on Computer Vision (ICCV)*, 2017.

[15] Meilu Liu, Yifan Xu, et al. RT-DETR: Real-time detection transformer. `https://arxiv.org/abs/2304.08069`, 2023.

[16] Ming Liu et al. Fast spatial pyramid pooling for object detection. *Neurocomputing*, 2021.

[17] Shu Liu et al. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016.

[19] Zhuang Liu et al. Convnext: A convnet for the 2020s. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[20] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] Joseph Redmon and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. `https://arxiv.org/abs/1804.02767`, 2018.

[24] Ultralytics. YOLOv8: Ultralytics next-generation object detector. `https://github.com/ultralytics/ultralytics`, 2023.

[25] Ultralytics. Ultralytics hub—cloud training and model management. `https://hub.ultralytics.com`, 2024.

[26] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Cspnet: A new backbone that can enhance learning capability of cnn. *CVPR Workshops*, 2020.

[27] Chien-Yao Wang, I-Hau Yeh, et al. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time detectors. https://arxiv.org/abs/2207.02696, 2022.

[28] Sonia Zagadouni et al. Multi-currency banknote classification using ensemble deep networks. *Applied Sciences*, 2021.

[29] Zhiqiang Zhao et al. MNAS-FPN: Learning latency-aware hierarchical neural architecture for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] Xizhou Zhu et al. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019.