Enhancing the Fairness and Performance of Edge Cameras with Explainable AI

Truong Thanh Hung Nguyen[†‡], Vo Thanh Khang Nguyen[‡], Quoc Hung Cao[‡],
Van Binh Truong[‡], Quoc Khanh Nguyen[‡], Hung Cao[†]
[‡]Quy Nhon AI, FPT Software, Vietnam [†]Analytics Everywhere Lab, University of New Brunswick, Canada
Email: {hungntt, khangnvt1, hungcq3, binhtv8, khanhnq33}@fpt.com, hcao3@unb.ca

2024 IEEE International Conference on Consumer Electronics (ICCE) | 979-8-3503-2413-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCE59016.2024.10444383

***Abstract*—The rising use of Artificial Intelligence (AI) in human detection on Edge camera systems has led to accurate but complex models, challenging to interpret and debug. Our research presents a diagnostic method using XAI for model debugging, with expert-driven problem identification and solution creation. Validated on the Bytetrack model in a real-world office Edge network, we found the training dataset as the main bias source and suggested model augmentation as a solution. Our approach helps identify model biases, essential for achieving fair and trustworthy models.**

***Index Terms*—Explainable AI, Edge Camera**

I. Introduction

Human detection through security cameras, a pivotal AI task, employs AI models like YOLO and its YOLOX variant for alerts, such as falls and intrusions. Specifically, Bytetrack, based on YOLOX, excels in multi-object tracking [1], [2]. Yet, it struggles in detecting obscured or disabled individuals (Fig. 1a, Fig. 1b). Given their black-box nature, these models pose debugging challenges. Though XAI aids debugging in tabular and text data [3], its use in image data is less explored. Hence, our paper introduces an XAI-driven framework to debug human detection models in security cameras. The approach leverages experts for diagnosing problems and proposing solutions, with potential wider relevance to object detection and classification.

(a) (b)

Fig. 1. (a) A security camera on the ceiling of an office can detect ordinary people (green boxes), but not people who cover their bodies with a cloth. (b) The Bytetrack model cannot detect the disabled woman but still detect the other, who is not disabled.

II. Related Work

A. Human Detection

Human detection identifies humans in images or videos and has evolved with various methods. Deep Learning (DL) brought forward models that address challenges like object size and illumination differences. Capitalizing on YOLOX's [1] success, Bytetrack [2] was designed for human detection, leveraging YOLOX for detection and Byte for post-processing.

B. Explainable AI

AI's integration into real-world scenarios has led to multiple Explainable AI (XAI) strategies: perturbation-based, backpropagation-based, and example-based. Perturbation techniques, such as D-RISE [4], which work independently of model design, perturb input images, then analyze predictions to gauge pixel or superpixel influence on outcomes. While widely applicable, their computational demand can be limiting. Backpropagation methods delve into model architecture to fetch explanatory data. Recognized techniques include Grad-CAM [5], SeCAM [6]. Example-based methods, like Influence Function [7], explain using training data samples to ascertain their effects on predictions. While XAI's application to object detection is complex due to the intricate models, some methods, such as D-RISE [4], D-CLOSE [8], and G-CAME [9], are adaptations from classification for object detection.

C. Debugging Model Framework with XAI

Many studies utilize XAI methods [10], primarily answering, *"Why does the model predict this?"* Yet, the follow-up, *"How can explanations improve the model?"* requires using XAI to better the AI system. No research has yet outlined a framework for debugging human detection models. This paper, therefore, introduces such a framework, leveraging XAI to pinpoint issues and improve model fairness and efficacy.

III. Methodology

We present a structured debugging model framework shown in Fig. 2, with seven sequential stages. Each stage relies on the results of its predecessor. Where multiple methods or assumptions exist per stage, we offer strategy selection guidelines. In this framework, XAI aids experts in identifying core model issues and suggesting performance-enhancing solutions.

A. Data Selection and Extraction of Predictions

Our framework starts by selecting a training dataset subset for model enhancement, addressing potential dataset concerns. Public datasets like CrowdHuman [11], used in Bytetrack training, can face data poisoning [12], affecting data quality and model results. Error detection in the model or dataset is optimized using random testing [13], which randomly picks

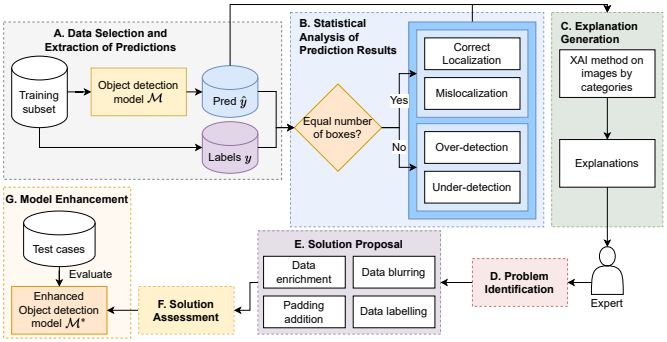

Fig. 2. The Debugging Framework for Human Detection Models

data for testing, spotting major flaws without full dataset checks. Based on the idea that small samples can be indicative, we use statistical sampling heuristics to set an optimal sample size, which should not surpass 10% of the full dataset or 1000 samples, ensuring a meaningful and efficient subset [14]. After selecting the data subset, it's fed into the model to generate predictions. These are then analyzed against the ground truth, helping gauge model metrics like accuracy, precision, and areas needing enhancement.

B. Statistical Analysis of Prediction Results

After obtaining predictions, they are categorized by comparing them with the ground-truth. This classification is guided by experts and, in our human detection context, results in four categories. Initially, dataset categorization relies on whether the model's predicted count aligns with the ground truth. Images are labeled as "Under-detection" if the model detects fewer people, and "Over-detection" if it detects more. If the model's count matches the ground truth, detection quality is evaluated by comparing model-detected boxes with ground truth boxes using Intersection over Union (IoU) values. Images with all box pairs having IoU ≥ 0.5 are deemed "Correct Localization", while others are "Mislocalization".

This process organizes the dataset based on prediction results, with three categories signaling potential model enhancements. The next stage delves deeper into error sources, laying the groundwork to boost the model's precision in detecting people within images.

C. Explanation Generation

In this phase, we use XAI methods to explain each image category. Given that D-RISE [4] is adaptable to diverse models without needing their architecture details and offers explanations for ground truth boxes (enabling comparison with model-detected boxes), we opt for D-RISE in human detection. These explanations assist experts in identifying the root of incorrect predictions in the following stage.

D. Problem Identification

Using the XAI results from the prior phase, experts analyze each category presented in the statistical analysis (Sec. III-B). The XAI indicates the model's focal regions on the input image. Experts assess these areas for relevance and potential biases. By comparing these regions across images in the same category, common patterns are identified. These patterns are then cross-referenced with other categories to spot shared features. Additionally, we compare XAI results across various models to further address potential challenges.

E. Solution Proposal

The solution proposal phase is important for enhancing model performance. Once the issue is identified, experts review the dataset and model to identify potential causes like data distribution, labels, biases, or model design. Solutions may involve tweaking model parameters, refining training data, or enhancing the training procedure.

F. Solution Assessment

Rather than implementing all possible solutions, we shall assess the feasibility of proposed solutions on a small dataset initially. We evaluate the advantages and disadvantages of each solution, drawing from prior case studies to assess their relevance to the present problem. The infeasible solutions can be identified and eliminated, thereby allowing for the selection of the most suitable solution.

G. Model Enhancement

After implementing the effective solution identified earlier, we refine the model to address issues highlighted in Sec. III-D. We then assess the model's enhancement by contrasting its performance pre and post-refinement, specifically comparing predictive metrics on initially selected images. Additionally, we might test using cases the original model struggled with to validate the model's enhanced capability in tackling the pinpointed issue.

IV. Experiment

In our study, we detail each step as illustrated in Fig. 2. We experiment using the Bytetrack model pre-trained on datasets like MOT17 [15], Cityperson [16], ETHZ [17], and CrowdHuman [11].

A. Data Selection and Prediction Extraction

Our training dataset amalgamates four public datasets [11], [15]–[17]. We use CrowdHuman for our tests, divided into training (15000 images), validation (4370 images), and testing (5000 images) sets. These sets, with a combined 470K human instances, offer varied bounding box annotations. We choose a random 1000-image subset from CrowdHuman's training set for extracting model predictions, as outlined in Sec. III-A.

B. Analyzing Prediction Results

Here, we match predicted boxes with the ground truth. "Under-detection" is the predominant issue, constituting 85.5%. While, "Under-detection" accounts for 17%, "Over-detection" accounts for 10.8%, and "Mislocalization" accounts for 20%.

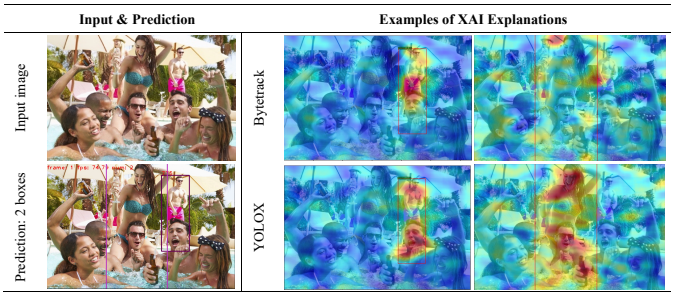

Fig. 3. Examples of XAI Explanations with Bytetrack and YOLOX model. In which, each image in the second column is the XAI Explanations for a corresponding box.

C. Explanation Generation

The Bytetrack model is a composite of YOLOX, responsible for detection, and the Byte phase that processes these detections. YOLOX is vital as the subsequent Byte step relies on its outputs. Byte's role is to maintain low-score predictions possibly hidden by other items [2]. We use D-RISE to interpret YOLOX, referencing the final box coordinates from Bytetrack [18]. Additionally, comparing Bytetrack and YOLOX using D-RISE on YOLOX's weights aids in identifying differences, showcased in Fig. 3 [1].

D. Problem Identification

The XAI explanations in Fig. 3 indicate Bytetrack's focus on entire human bodies, exposing its struggle to detect individuals showing only their heads. Experiments with images of people in wheelchairs, where bodies are partly concealed, amplify this limitation, with the model overlooking them as seen in Fig. 1b. Similar misses happen with people hidden behind objects, highlighted in Fig. 1a. Hence, Bytetrack's challenge in spotting partially visible humans emerges as a key concern needing attention and resolution.

E. Solution Proposal

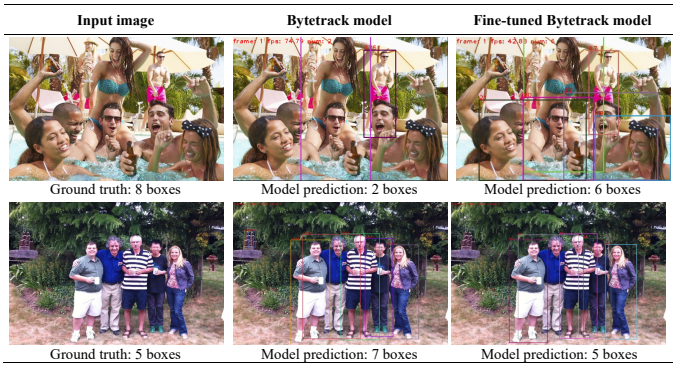

Fig. 4. Predictions of the Bytetrack model before and after fine-tuning.

We pinpointed specific issues and proposed assumptions accordingly:

- Dataset: On average, images have 23 people, making heads smaller than bodies, potentially leading to a body bias. We also suspect label issues with ground truth box coordinates outside the image, shown in Fig. 3 and Table II.
- Model: Bytetrack tries to resolve occluded objects [2]. For head-only images, Bytetrack expects an associated body.

TABLE I
GROUND TRUTH BOXES' COORDINATE OF THE INPUT IMAGE IN THE FIRST ROW OF FIG. 3, WHERE 7/8 BOXES ARE OUTSIDE THE IMAGE.

Left	-50	-12	308	499	618	608	318	303
Top	35	87	292	171	370	61	-14	-3
Right	531	451	635	988	1034	758	673	444
Bottom	131	1325	1228	1201	1243	444	745	437
Outside image	×	×	×	×	×		×	×

Proposed solutions include:

- Data enrichment: Add images with mostly obscured body sections.
- Data blurring: Based on XAI findings, blur bodies to make the model focus on heads.
- Padding: Ensure bounding boxes are fully within images.
- Relabeling: Adjust bounding boxes to remain inside the image.

F. Solution Assessment

We conduct a comprehensive analysis to identify and implement the most suitable solution to the problem. Each solution is evaluated as follows:

- Data enrichment: The current dataset already has partly hidden figures, so more data might not help much.
- Data blurring: Effective for image classification, but might not suit human detection where only humans are predicted.
- Padding: While sometimes effective, as in Fig. 5, it often fails, especially when objects obstruct people.
- Relabeling: Given dataset inconsistencies and variant model features, relabeling seems promising.

Following this analysis, relabeling emerges as the most impactful solution.

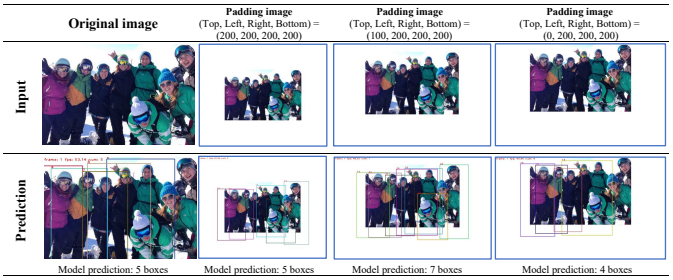

Fig. 5. Example of padding result. (Top, Left, Right, Bottom) = (100, 200, 200, 200) signifies padding of 100, 200, 200, and 200 pixels respectively on the top, left, right, and bottom.

G. Model and Dataset Enhancement

The CrowdHuman dataset is reannotated by constraining bounding box coordinates within the image dimensions, as delineated by $x'_{\text{top, left}} = \texttt{max}(0, x_{\text{top, left}})$, $y'_{\text{top, left}} = \texttt{max}(0, y_{\text{top, left}})$, $x'_{\text{bottom, right}} = \texttt{min}(w, x_{\text{bottom, right}})$,

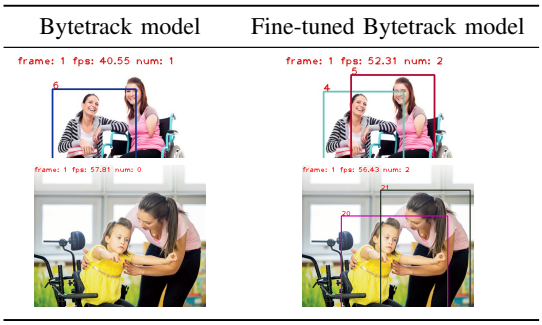

Fig. 6. Model's prediction on physically disabled person images. After fine-tuning, the model performs better than the original pre-trained model.

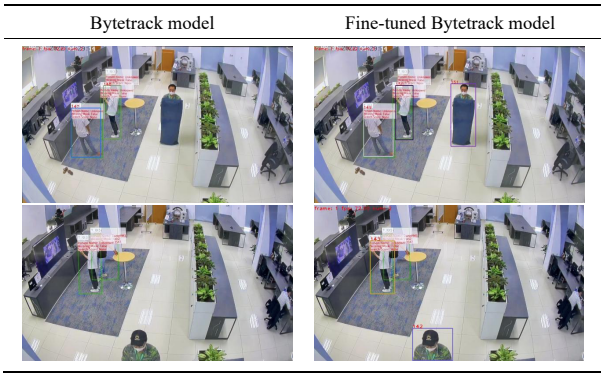

Fig. 7. Model's prediction on a security camera. The fine-tuned model performs better than the original pre-trained model detecting covered people.

$y'_{\text{bottom, right}} = \min(h, y_{\text{bottom, right}})$. Here, w, h represents the image's width and height, respectively. The coordinates $(x'_{\text{top, left}}, y'_{\text{top, left}})$ and $(x'_{\text{bottom, right}}, y'_{\text{bottom, right}})$ denote the adjusted top-left and bottom-right points, respectively. Subsequent model refinement occurs over 10 epochs, with performance enhancement evaluated in three scenarios:

- Training Dataset Testing: We test a 1000-image subset after refining the model. Both quantitative and qualitative evaluations are made against the original model, as seen in Table II and Fig. 4. The updated model better localizes in 855 "Under-detection" images, improving by 21 cases.
- Images of Disabled Individuals: The adjusted model shows better detection in images featuring physically disabled people, highlighted in Fig. 6.
- Detection in Surveillance Footage: We assess the model in real-life contexts, like office security footage where people might be partly hidden. Post-refinement performance, showcasing improvements, is depicted in Fig. 7.

TABLE II
STATISTICAL RESULT PRE-TRAINED MODEL VERSUS FINE-TUNED MODEL. THE ARROW ↑/↓ INDICATES THE HIGHER/LOWER VALUE, THE BETTER. THE BOLD INDICATES THE BETTER RESULT.

Case	Pre-trained model	Fine-tuned model
Under-detection (↓)	855	**834**
Over-detection (↓)	17	**13**
Correct Localization (↑)	108	**133**
Mislocalization (↓)	20	20

V. Conclusion and Future Work

This study introduces a human detection debugging framework using XAI aided by experts. Our approach pinpoints data labeling as a significant issue in Bytetrack's biases and can adapt to other detection problems, especially those focusing on specific classes.

References

[1] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: exceeding YOLO series in 2021," *CoRR*, vol. abs/2107.08430, 2021. [Online]. Available: https://arxiv.org/abs/2107.08430

[2] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," *CoRR*, vol. abs/2110.06864, 2021. [Online]. Available: https://arxiv.org/abs/2110.06864

[3] R. Yousefzadeh and D. P. O'Leary, "Auditing and debugging deep learning models via decision boundaries: Individual-level and group-level analysis," *CoRR*, vol. abs/2001.00682, 2020. [Online]. Available: http://arxiv.org/abs/2001.00682

[4] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. Saenko, "Black-box explanation of object detectors via saliency maps," *CoRR*, vol. abs/2006.03204, 2020. [Online]. Available: https://arxiv.org/abs/2006.03204

[5] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016. [Online]. Available: http://arxiv.org/abs/1610.02391

[6] P. Nguyen, H. CAO, K. NGUYEN, H. NGUYEN, and T. YAIRI, "Secam: Tightly accelerate the image explanation via region-based segmentation," *IEICE Transactions on Information and Systems*, vol. E105.D, pp. 1401–1417, 08 2022.

[7] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," 2017. [Online]. Available: https://arxiv.org/abs/1703.04730

[8] V. B. Truong, T. T. H. Nguyen, V. T. K. Nguyen, Q. K. Nguyen, and Q. H. Cao, "Towards better explanations for object detection," *arXiv preprint arXiv:2306.02744*, 2023.

[9] Q. K. Nguyen, T. T. H. Nguyen, V. T. K. Nguyen, V. B. Truong, and Q. H. Cao, "G-came: Gaussian-class activation mapping explainer for object detectors," *arXiv preprint arXiv:2306.03400*, 2023.

[10] T. T. H. Nguyen, V. B. Truong, V. T. K. Nguyen, Q. H. Cao, and Q. K. Nguyen, "Towards trust of explainable ai in thyroid nodule diagnosis," *arXiv preprint arXiv:2303.04731*, 2023.

[11] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *CoRR*, vol. abs/1805.00123, 2018. [Online]. Available: http://arxiv.org/abs/1805.00123

[12] R. S. S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia, "Adversarial machine learning - industry perspectives," *CoRR*, vol. abs/2002.05646, 2020. [Online]. Available: https://arxiv.org/abs/2002.05646

[13] J. Mayer and C. Schneckenburger, "An empirical analysis and comparison of random testing techniques," in *Proceedings of the 2006 ACM/IEEE international symposium on Empirical software engineering*, 2006, pp. 105–114.

[14] C. R. W. VanVoorhis and B. L. Morgan, "Understanding power and rules of thumb for determining sample sizes," 2007.

[15] A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *CoRR*, vol. abs/1603.00831, 2016. [Online]. Available: http://arxiv.org/abs/1603.00831

[16] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," *CoRR*, vol. abs/1702.05693, 2017. [Online]. Available: http://arxiv.org/abs/1702.05693

[17] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[18] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. Saenko, "Black-box explanation of object detectors via saliency maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 443–11 452.