Towards Robust Object Detection in Unseen Catheterization Laboratories

Wang, Zipeng ; Butler, Rick; van den Dobbelsteen, John; Hendriks, Benno; van der Elst, Maarten; Dauwels, Justin

DOI
10.1109/MeMeA60663.2024.10596906

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the 2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)

2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA) | 979-8-3503-0799-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/MEMEA60663.2024.10596906

Towards Robust Object Detection in Unseen Catheterization Laboratories

Zipeng Wang[1], Rick Butler[1], John J. van den Dobbelsteen[1], Benno H. W. Hendriks[1,2], Maarten Van der Elst[1,3], and Justin Dauwels[1*]

[1]Delft University of Technology, Delft, the Netherlands
[2]Philips Research Laboratories, Eindhoven, the Netherlands
[3]Reinier de Graaf Group, Delft, the Netherlands
[*]Corresponding Author (J.H.G.Dauwels@tudelft.nl)

***Abstract*—Deep learning-based object detectors, while offering exceptional performance, are data-dependent and can suffer from generalization issues. In this work, we investigated deep neural networks for detecting people and medical instruments for the vision-based workflow analysis system inside Catheterization Laboratories (Cath Labs). The central problem explored in this paper is the fact that the performance of the detector can degrade drastically if it is trained and tested on data from different Cath Labs. Our research aimed to investigate the underlying causes of this specific performance degradation and find solutions to mitigate this issue. We employed the YOLOv8 object detector and created datasets from clinical procedures recorded at Reinier de Graaf Hospital (RdGG) and Philips Best Campus, supplemented with publicly accessible images. Through a series of experiments complemented by data visualization, we discovered that the performance degradation primarily stems from data distribution shifts in the feature space. Notably, the object detector trained on non-sensitive online images can generalize to unseen Cath Labs, outperforming the model trained on a procedure recording from a different Cath Lab. The detector trained on the online images achieved an mAP@0.5 of 0.517 on the RdGG dataset. Furthermore, by switching to the most suitable camera for each object in the Cath Lab, the multi-camera system can further improve the detection performance significantly. An aggregated 1-camera mAP@0.5 of 0.679 is achieved for single-object classes on the RdGG dataset.**

***Index Terms*—Object Detection, Catheterization Laboratory, Domain Shift, Clinical Workflow Analysis**

I. Introduction

A Catheterization Laboratory (Cath Lab) is a specialized procedural room in hospitals, equipped with medical imaging instruments to visualize heart chambers and vessels [1]. It is essential for the diagnosis and treatment of cardiovascular diseases. For example, Diagnostic Cardiac Catheterization requires a cardiologist to insert a catheter through an artery and finally into the heart via the guidance of a medical imaging instrument to find blockages or narrowings [2]. However, various threats and risks for both medical personnel and patients are associated with Cath Lab. For instance, Chronic radiation exposure can pose health concerns for interventional physicians, despite protective measures like lead aprons [3].

To provide insights for efficiency improvement and risk minimization, a vision-based measurement system has been designed in earlier works for workflow analysis. The system has been deployed inside Reinier de Graaf Hospital in Delft and the Philips Best Campus in Eindhoven. It consists of multiple cameras surrounding the operating table to provide a comprehensive view of the procedures. Variable of interest can be measured by running video analysis algorithms, e.g. the distance between medical personnel and the X-ray machine for studying X-ray exposure risks. A deep learning-based object detector plays a central role in video analysis by identifying and locating medical staff and instruments within images.

Prior research has demonstrated the effectiveness of measurement systems featuring multi-camera setups paired with deep learning-based models for measuring the pose and location of clinicians [4], [5]. For optimal performance, researchers trained their machine learning model on data from the same test environment. However, the reliance on deep learning introduces challenges in generalizing to new environments. Research on machine learning models suggests they tend to provide erroneous predictions when encountering data that follow different distributions than the training data [6]. When we expand our measurement system to previously unseen Cath Labs, the YOLOv8 object detector [7] fails to give reliable results. The traditional solution is acquiring and annotating data inside the new Cath Lab, followed by retraining and the model. However, as access to sensitive medical data is highly restricted, this practice makes deploying the measurement system time-consuming and expensive.

Therefore, the primary aim of this study is to explore the causes behind the performance degradation of the YOLOv8 object detector in unseen Cath Labs and to identify strategies to overcome this challenge. This investigation lays the groundwork for creating vision-based measurement systems that are robust for effective deployment in previously unseen Cath Labs. Our main contributions are as follows:

1) We collected clinical data from two Cath Labs and demonstrated that detection performance degradation occurs when the object detector processes images from unseen Cath Labs. Additionally, We further demonstrated the performance drop stems from divergent data distributions in the feature space, which provides insights

for further research on the generalization ability of the measurement system.
2) We illustrated that leveraging publicly available online images as alternative training data, coupled with utilizing video data from multiple views, can improve the generalization ability of the detector to unseen Cath Labs. They are crucial strategies for developing robust vision-based measurement systems for clinical purposes.

II. RELATED WORKS

A. Object Detection

Object detection stands as a cornerstone task in computer vision, critical for understanding images and videos [8]. Generic object detection, the most prevalent form, involves identifying and locating objects within an image from predefined categories [9]. Models accomplish this by assigning a bounding box and confidence score to each detected object.

The introduction of Region-based Convolutional Neural Networks (R-CNN) in 2014 marked a pivotal shift in object detection research towards deep-learning-based methods, which can be broadly divided into two directions, one-stage and two-stage detectors [8]. The two-stage detectors, featuring R-CNN, first propose regions of interest in the input image, and each region is subsequently classified into one of the predefined categories by a Convolutional Neural Network (CNN) [10].

One-stage object detectors operate in a more unified manner. Those detectors treat object detection as a regression problem, and predict bounding boxes and corresponding class probabilities [11]. One-stage object detectors first extract feature maps from the input image. Next, regression is performed on each cell of the feature maps to predict bounding boxes and their corresponding class probabilities. The design paradigm of one-stage object detectors is established by You Only Look Once (YOLO) [11]. One-stage object detectors tend to deliver better speed and generalization ability than two-stage object detectors, although their localization accuracy and detection performance are suboptimal for small objects in the first-generation YOLO. In later generations of YOLO, these issues have been addressed. For instance, YOLOv3 improves the detection performance of small-scale objects with a network structure resembling the Feature Pyramid Network [12]. As of this study, YOLOv8 represents the latest advancement in the YOLO series [7]. It utilizes a large combination of techniques compared to its predecessor and has achieved an mAP@0.5-0.95 of 0.539 on the 2017 COCO val dataset [7]. Therefore, we chose YOLOv8 as the object detector in this work.

B. Domain Shift

Domain shifts refer to the domain-related data distribution difference, which can damage the performance of machine learning methods [6]. A common assumption in machine learning is that both training and testing data are sampled from the same distribution, a condition that often does not reflect reality [13]. In practical applications, training data from the source domain and testing data from the target domain can have distribution differences. Mathematically speaking, there are different kinds of domain shifts: covariate shift, label shift, and concept shift, among other more general data distribution shifts [14].

A machine learning model leverages input features, denoted by X, to predict target variables, represented as Y. This process can be achieved by estimating the conditional probability $P(Y|X)$. Different types of domain shifts can be depicted by the change in decomposed components of the joint distribution, expressed as $P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$.

1) Covariate shift assumes $P_{\text{train}}(X) \neq P_{\text{test}}(X)$, $P_{\text{train}}(Y|X) = P_{\text{test}}(Y|X)$
2) Label shift assumes $P_{\text{train}}(Y) \neq P_{\text{test}}(Y)$, $P_{\text{train}}(X|Y) = P_{\text{test}}(X|Y)$
3) Concept shift assumes $P_{\text{train}}(Y|X) \neq P_{\text{test}}(Y|X)$, $P_{\text{train}}(X) = P_{\text{test}}(X)$

Our case fits the covariate shift assumption, where the object appearances from different Cath Labs are only tiny subsets of the entire spectrum of possible images, while the concept of the object class (the relationship between the RGB image and its object class) remains constant. Although $P(Y|X)$ is unchanged, covariate shift will make it difficult to estimate $P(Y|X)$ in regions where the training data points are sparse or absent.

Domain shifts arise from a variety of factors and pose serious threats to machine learning systems deployed in the real world. Researchers have found its presence in images taken by different types of cameras [15]. In the medical field, different imaging devices can also impair the performance of detection systems, e.g., polyp detection in the digestive system [16]. The list of factors that cause domain shift can be nearly unlimited. Most commonly, they are changes in lighting, camera angles, or backgrounds [17]. The ubiquitous nature of domain shift makes it difficult to avoid. This issue is particularly critical in high-stakes domains such as autonomous driving and healthcare, where the consequences of errors can be lethal, thus research on domain shift has attracted growing interest [18].

III. METHODS

A. Data Distribution Visualization

The literature on domain shifts highlights how shifts in data distribution can lead to performance degradation in machine learning models. This insight drives our investigation into whether the images in our datasets follow distinct distributions in the feature space. For this purpose, we have applied visualization methods to show the feature distribution of the images in our datasets. Similar visualization methods have been applied in research regarding domain shift [19].

One-stage object detectors, including YOLOv8, first extract features from the image and then perform regression to obtain predictions from the feature maps. Our visualization method aims to show that the data distribution shifts occur in the feature maps, which is highly problematic for the following regression task. For simplicity, in this project, we chose the

deepest layer preceding the detection head in YOLOv8 to obtain the feature maps. As the feature maps have high dimensionality ($512 \times 20 \times 12$), we need to further reduce them to a 2-dimensional vector for visualization.

The first step is calculating feature channel statistics as a feature vector. The feature channel statistics capture the activation patterns of the neural networks. We calculate the mean and variance of each layer of the feature maps as the feature channel statistics. The mean provides a measure of the average intensity of the activations, and variance measures how much the activations vary across the feature map.

In addition to calculating the feature vector of the whole image, we also calculate the feature vector on the object level. We can read the position of an object instance from the annotation, and discretize it to get the responsible cell in the feature maps. The receptive field of the head of YOLOv8 is 5×5. Therefore, we can calculate this 5×5 region around the cell, rather than the whole feature maps, to get the feature vector on the object level.

The second step is reducing the feature vector to 2 dimensions. It is achieved by applying dimensionality reduction methods on the feature vector, whose size is originally 1024. We have experimented with Principal Component Analysis (PCA) [20], T-distributed Stochastic Neighbor Embedding (T-SNE) [21], and Uniform Manifold Approximation and Projection (UMAP) [22]. We chose UMAP as the dimensionality reduction method, as it can preserve both local and global data structures [22]. After obtaining the reduced feature vector of each image (or object), we plotted each of them as a point to demonstrate the distribution of our datasets.

B. Multi-camera System Evaluation for Camera Switching

Multi-camera system evaluation can illustrate their effectiveness in improving detection performance by addressing challenges associated with viewpoint changes and occlusions. Objects seen from a significantly different viewpoint than in the training data or highly occluded are difficult to detect and tend to have low detection confidence. For this issue, the redundancy offered by a multi-camera system can be leveraged. Camera switching is performed on every single-object class for simplicity. For each object and in every frame, we select top-N camera views with the highest detection confidence score, where N can be 1 or 2. For tasks that require only the state of the object, a single camera is sufficient. An example of such tasks is fall detection [23]. Conversely, tasks demanding 3D information about an object, such as 3D pose estimation, require at least two cameras [4]. For evaluation purposes, we generate an aggregated version of the object detection metrics, based on this camera switching strategy.

Commonly used metrics for the object detection task, such as Average Precision (AP) and mean Average Precision (mAP), are derived from precision and recall [24]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\text{All detections}}, \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{All ground truths}}, \tag{2}$$

where TP, FN, and FP are the True Positive, False Negative, and False Positive, respectively.

Aggregated metrics are evaluated per multi-view frame. For each frame, an object is examined if it is detectable, detected, and correctly detected. 'Detectable' suggests the object is visible in at least N camera views. 'Detected' means in the N selected camera views, the detection confidence is higher than the confidence threshold. We consider an object 'correctly detected', if both the detection confidence and Intersection over Union (IoU) are higher than the thresholds in the N selected camera views. We count those frames for calculating aggregated metrics.

After substituting the corresponding elements in precision and recall, aggregated precision and aggregated recall are obtained. Aggregated AP are derived from them.

$$\text{Precision}_{\text{aggregated}} = \frac{N_{\text{correctly detected}}}{N_{\text{detected}}}, \tag{3}$$

$$\text{Recall}_{\text{aggregated}} = \frac{N_{\text{correctly detected}}}{N_{\text{detectable}}}, \tag{4}$$

where $N_{\text{detectable}}$, N_{detected}, and $N_{\text{correctly detected}}$ are the number of multi-view frames that fit our evaluation definition.

IV. EXPERIMENT AND RESULT

A. Datasets

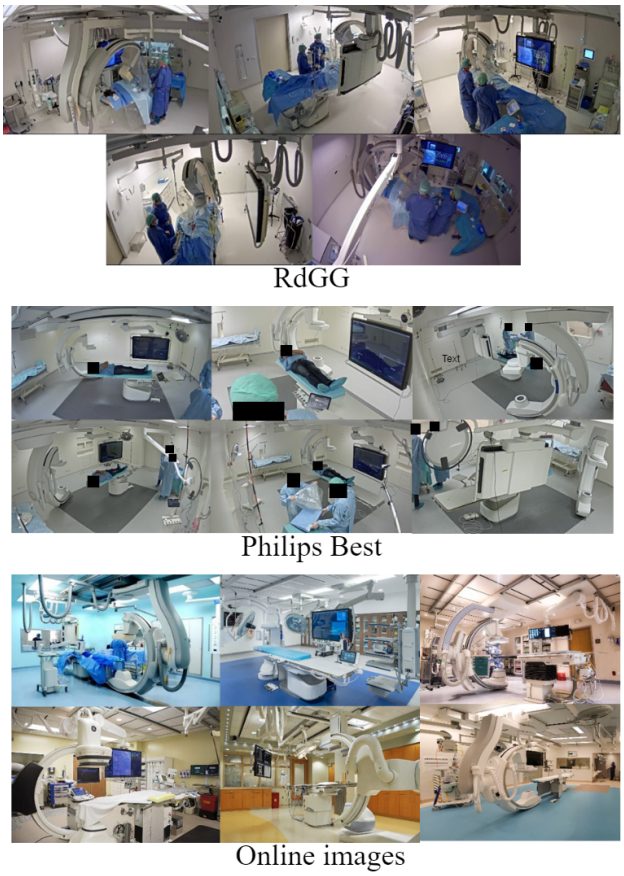


Fig. 1: Example images of our datasets.

In this project, we have clinical procedure datasets collected inside Cath Labs and an online image dataset collected from online search engines as alternative training data. The object classes for detection are doctor, patient, operating table, instrument table, control panel display, control panel button, x-ray detector, x-ray source, and display.

Procedure datasets include a dataset collected at the Reinier de Graaf Hospital, alongside two datasets from Philips Best Campus. The RdGG dataset comprises images from a real procedure, recorded by a 5-camera system operating at a frame rate of 25 Frames Per Second (FPS), totaling 3100 images. Conversely, the Philips Best 1 and 2 datasets contain 492 and 792 images respectively, and are collected from mock procedures. These were recorded using a similar 6-camera system at 25 FPS. In all three clinical datasets, images are extracted at 5-second intervals from the video recordings. Since the images from each clinical dataset are captured inside a specific Cath Lab, the images have limited variability in the objects and background across the dataset (intra-variability). However, the objects and backgrounds from different Cath Labs vary substantially (inter-variability).

The online image dataset, sourced from Google and Bing using the keyword 'Catheterization Laboratory', has a wide variety of objects and background appearances. It contains 800 images of different Cath Labs. However, most of the images are captured for commercial purposes with a highly limited viewpoint variety and few occlusions.

B. Experiment Design

Our experiments are designed to show the relation between the performance gap and data distribution shifts, along with the effectiveness of the mitigating solutions.

The first experiment aims to show the performance gap and its related distribution shifts. We chose Philips Best 1 as the testing set. Philips Best 2 dataset, RdGG dataset, and the online image dataset were chosen as the training set respectively. These three combinations simulate three scenarios: Training and testing in the same Cath Lab, training and testing in different Cath Labs, and the detector trained on public-available data. To gain insights in the experimental results, we performed data visualization.

The second experiment is designed to investigate the detector's generalization ability when it is only trained on publicly available data. We chose the online images dataset as the training data, and the three clinical datasets as the test set respectively. Additionally, we visualized the object-level data distribution of the best-performing class and the worst-performing class. It visualizes whether the object detection performance of each class is related to the object-level data distribution.

The third experiment is designed to assess the multi-camera system. We adapt the camera switching strategies while keeping the same training set and test set as in the second experiment. Aggregated metrics of multi-camera detection are used for evaluation and compared to object detection metrics for single-camera detection.

C. Results and Analysis

TABLE I: Evaluation results (AP@0.5) of the YOLOv8 object detector on the Philips Best 1 dataset, when the detector is trained on different datasets.

Class	Philips Best 2	RdGG	Online images
Doctor	0.942	0.778	0.831
Patient	0.928	0.118	0.516
Operating table	0.929	0.404	0.581
Instrument table	0.830	0.240	0.670
Control panel display	0.849	0.183	0.214
Control panel button	0.930	0.059	0.054
X-ray detector	0.976	0.012	0.766
X-ray source	0.853	0.047	0.630
Display	1.000	0.775	1.000
Mean	0.915	0.291	0.585

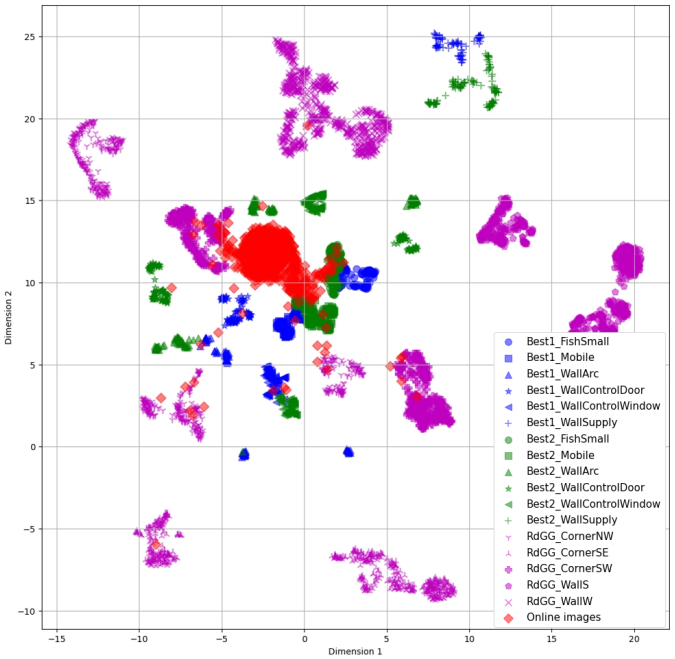
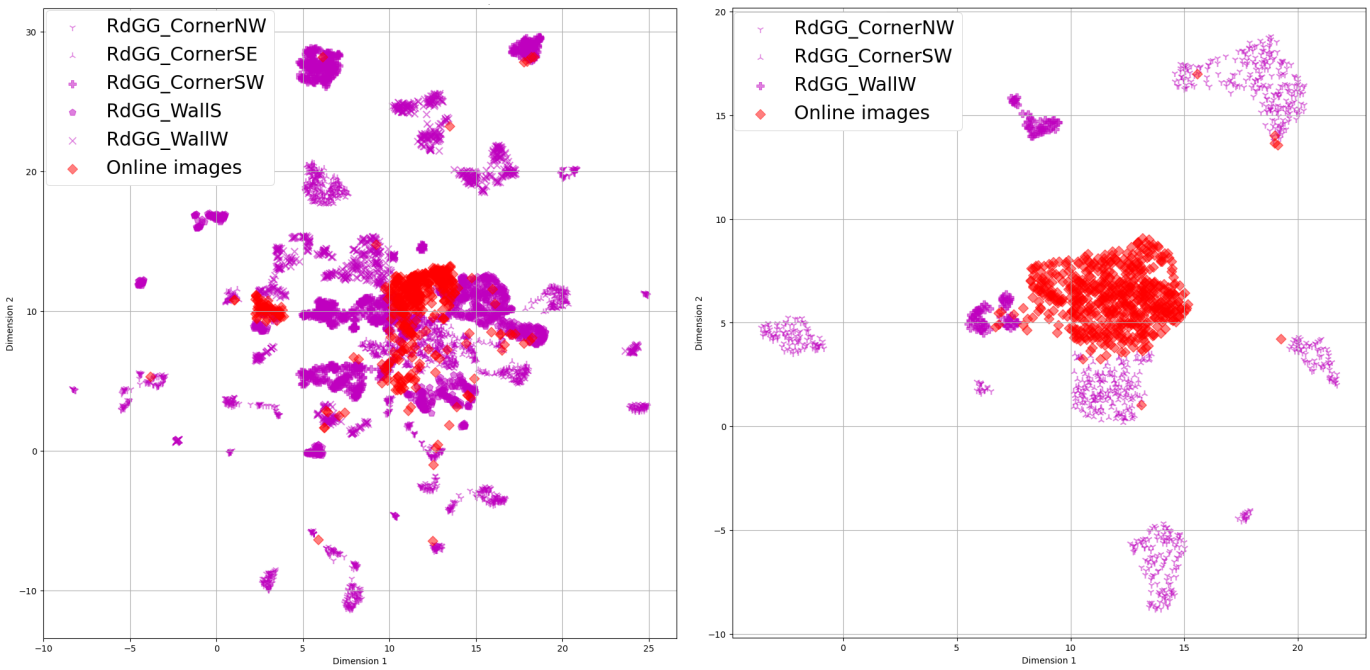


Fig. 2: UMAP data distribution visualization of our datasets. (Markers denote camera view, purple represents the RdGG dataset, blue and green represent the Philips Best 1 and 2 datasets respectively, and red represents the online image dataset.)

Table I presents the results of the first experiment. Results from the first two columns suggest that training the detector in the same Cath Lab significantly outperforms training it in a different Cath Lab. The performance of the model trained on online images falls in between, achieving an mAP@0.5 of 0.585. Data visualization shown in Fig. 2 provides an explanation for this performance gap. Images from the same Cath Labs have relatively close data distributions, while images from different Cath Labs have divergent data distributions. The online images, compared to clinical procedure recordings, have a wider data distribution, covering more testing data than images in the RdGG dataset.



Doctor, AP@0.5=0.856 X-ray source, AP@0.5=0.071

Fig. 3: UMAP object-level data visualization of the best and worst performing class when the model is trained on the online image dataset and tested on the RdGG dataset. (Purple represents the RdGG dataset, and red represents the online image dataset.)

TABLE II: Evaluation results (AP@0.5) of the YOLOv8 object detector on different clinical procedure datasets, when the detector is trained only on the online image dataset.

Class	Philips Best 1	Philips Best 2	RdGG
Doctor	0.831	0.870	0.856
Patient	0.516	0.485	0.274
Operating table	0.581	0.649	0.704
Instrument table	0.670	0.616	0.266
Control panel display	0.214	0.152	0.738
Control panel button	0.054	0.115	0.311
X-ray detector	0.766	0.684	0.709
X-ray source	0.630	0.681	0.071
Display	1.000	0.974	0.727
Mean	0.585	0.581	0.517

According to the results presented in Table II, the object detector trained purely on online images can generalize to previously unseen Cath Labs with a moderately good performance. However, most object classes have inconsistent detection performance across different Cath Labs. Fig. 3 shows the object-level data distribution of the best and worst performing class when the model is trained on online images and tested on the RdGG dataset. The doctors can be reliably detected in the different datasets. We noticed that the data distributions for doctors in the training and test sets are similar. Therefore, good generalization across different datasets can indeed be expected. However, the X-ray source class has very distinct data distribution in the training set and test set, which may explain the poor detection results.

Table III shows the aggregated average precision when we apply the camera switching strategy. Compared to the results in Table II, the multi-camera system can better detect objects by switching to the most confident (mostly with the least occlusion and viewpoint changes) camera view.

D. Discussion

Our results demonstrated the weakness in the generalization ability of the vision-based measurement system, which is the object detector of the system delivering poor detection performance in unseen Cath Labs, achieving an mAP@0.5 of 0.291 in our experiments. However, the issue can be mitigated by introducing publicly available images for training data diversity and incorporating multiple camera views. These two strategies lay a promising foundation for developing robust, vision-based measurement systems for deployment in unseen Cath Labs.

However, the results also highlight concerns regarding varying detection performance in unseen Cath Labs, which calls for more safety measures for system deployment. Data visualization shows images from different Cath Labs following distinct data distributions. Deploying the object detector in an unseen Cath Lab introduces uncertainty in data distribution compared to our training data, potentially leading to poor detection performance. On the bright side, we have illustrated divergent data distribution causing object detection performance degradation in unseen Cath Lab. It motivates for incorporation of an Out-of-Distribution (OoD) detector inside the measurement system to alert for potential detection inaccuracies when encountering unfamiliar objects. In the

TABLE III: Aggregated AP@0.5 of the YOLOv8 object detector trained on the online image dataset and evaluated on different clinical procedure datasets when using the most confident 1 or 2 camera(s).

	Aggregated 1-camera AP@0.5			Aggregated 2-camera AP@0.5		
	Philips Best 1	Philips Best 2	RdGG	Philips Best 1	Philips Best 2	RdGG
Patient	0.947	0.832	0.505	0.570	0.459	0.053
Operating table	0.884	0.934	0.958	0.764	0.827	0.803
Instrument table	0.896	0.793	0.585	0.809	0.496	0.076
Control panel display	0.440	0.334	0.897	0.105	0.031	0.596
Control panel button	0.041	0.211	0.635	0.012	0.026	0.227
X-ray detector	0.956	0.986	0.798	0.918	0.946	0.605
X-ray source	1.000	1.000	0.053	0.908	0.790	0
Display	1.000	1.000	1.000	1.000	1.000	1.000
Mean	0.771	0.761	0.679	0.636	0.572	0.420

medical field, OoD detection has drawn increasing interest as both benchmarks and OoD detectors have been proposed by researchers [25], [26].

V. Conclusion

Our study demonstrates how diverse training data and multi-camera systems significantly enhance the detection of medical personnel and instruments in previously unseen Cath Labs. For this purpose, we employed the YOLOv8 object detector and created datasets from clinical procedures recorded at Reinier de Graaf Hospital and Philips Best Campus, supplemented with publicly accessible images. Through a series of experiments complemented by data visualization, we discovered that the performance degradation primarily stems from data distribution shifts in the feature space. By collecting diverse training data and adapting a camera switching strategy for the multi-camera system, we can alleviate the distribution difference in the training and inference data to achieve a more robust object detection performance in unseen Cath Labs.

References

[1] C. E. Chambers, K. A. Fetterly, R. Holzer, P.-J. Paul Lin, J. C. Blankenship, S. Balter, and W. K. Laskey, "Radiation safety program for the cardiac catheterization laboratory," *Catheterization and Cardiovascular Interventions*, vol. 77, no. 4, pp. 546–556, 2011.

[2] R. A. Lange and L. D. Hillis, "Diagnostic cardiac catheterization," *Circulation*, vol. 107, no. 17, pp. e111–e113, 2003.

[3] E. Picano, M. G. Andreassi, E. Piccaluga, A. Cremonesi, and G. Guagliumi, "Occupational risks of chronic low dose radiation exposure in cardiac catheterisation laboratory: the italian healthy cath lab study," *EMJ Int Cardiol*, vol. 1, no. 1, pp. 50–8, 2013.

[4] V. Belagiannis, X. Wang, H. B. B. Shitrit, K. Hashimoto, R. Stauder, Y. Aoki, M. Kranzfelder, A. Schneider, P. Fua, S. Ilic *et al.*, "Parsing human skeletons in an operating room," *Machine Vision and Applications*, vol. 27, pp. 1035–1046, 2016.

[5] V. F. Rodrigues, R. S. Antunes, L. A. Seewald, R. Bazo, E. S. dos Reis, U. J. dos Santos, R. d. R. Righi, L. G. d. S. Junior, C. A. da Costa, F. L. Bertollo *et al.*, "A multi-sensor architecture combining human pose estimation and real-time location systems for workflow monitoring on hybrid operating suites," *Future Generation Computer Systems*, vol. 135, pp. 283–298, 2022.

[6] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[7] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[8] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.

[9] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, pp. 261–318, 2020.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[12] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[13] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2008.

[14] W. M. Kouw and M. Loog, "An introduction to domain adaptation and transfer learning," *arXiv preprint arXiv:1812.11806*, 2018.

[15] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer, 2010, pp. 213–226.

[16] X. Liu and Y. Yuan, "A source-free domain adaptive polyp detection framework with style diversification flow," *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1897–1908, 2022.

[17] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, 2011, pp. 1521–1528.

[18] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," *arXiv preprint arXiv:2108.13624*, 2021.

[19] Q. Fan, M. Segu, Y.-W. Tai, F. Yu, C.-K. Tang, B. Schiele, and D. Dai, "Towards robust object detection invariant to real-world domain shifts," in *The Eleventh International Conference on Learning Representations*, 2022.

[20] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[21] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[22] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction. arxiv 2018," *arXiv preprint arXiv:1802.03426*, 1802.

[23] S. Ezatzadeh, M. R. Keyvanpour, and S. V. Shojaedini, "A human fall detection framework based on multi-camera fusion," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 34, no. 6, pp. 905–924, 2022.

[24] R. Padilla, S. L. Netto, and E. A. Da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 international conference on systems, signals and image processing (IWSSIP)*. IEEE, 2020, pp. 237–242.

[25] T. Cao, C.-W. Huang, D. Y.-T. Hui, and J. P. Cohen, "A benchmark of medical out of distribution detection," *arXiv preprint arXiv:2007.04250*, 2020.

[26] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," *Advances in neural information processing systems*, vol. 32, 2019.