Atomic Spectroscopy

Article

Smart Surveillance for sustainable cities: Real-time missing child retrieval framework for video surveillance system

K. Iyshwarya Ratthi, *B. Yogameena
Research Scholar, Department of Electronics and Communication Engineering, Thiagarajar College of Engineering, Thiruparankundram, Madurai,Tamil Nadu, India
Professor, Department of Electronics and Communication Engineering, Thiagarajar College of Engineering, Thiruparankundram, Madurai, Tamil Nadu, India

ABSTRACT: A significant portion of law enforcement resources are dedicated to responding to reports of missing children. Video surveillance systems enhance the security and surveillance of public places such as airports, railway stations, and shopping malls. Despite this, the vast amount of video footage available can make it difficult to locate a lost child. In crowded areas and in a wide variety of settings, the problem is particularly acute. Firstly, we developed a dataset and a model for retrieving missing children. We present the first ever SmartChildren dataset derived from CCTV footage, comprising 8000 images spanning several ages and environments. Datasets are annotated as boy child, girl child and an adult. It is considered appropriate to label individuals under the age of twelve as children, and individuals over the age of twelve as adults. Based on the baseline of You Only Look Once version 5, the proposed YOLOv5 ChildLocate System detects individuals and labels them as girl, boy and an adult. Upon receiving a query for a missing girl child, the retrieval model will return that child from the gallery set of detected girl children who match the query. Hence, retrieval time is reduced. The Child Locate YOLOv5 works well with an 86.4 (1.2) % increase in mean Average Precision with 52fps compared to YOLOv5 with 85.2% with 45fps. And the child-adult retrieval network with mAP of 86.56% compared with MSCAN with 86.24%. A comparative analysis of the model is conducted with the developed dataset using six deep learning methods: ResNet50, Faster RCNN, R-FCN, SSD, YOLOv3, and YOLOv5. A variety of lighting and environmental conditions are used in experiments and ablation studies.

Keywords: Video surveillance, Deep learning, Missing Children Detection, Missing Child Retrieval, YOLOv5 ChildLocate System, Artificial Intelligence, Neural networks.

INTRODUCTION

The Indian population is over half young and under 18, with over 400 million youngsters under 18 living there [1]. There are several possible scenarios in which a child who has gone missing in one area of the country may be found in another area or state. Moreover, lost child retrieval has not been adequately researched. Therefore, it is challenging to identify a child among the reported abduction instances. It is still impossible to determine the exact number of abducted or traced children today due to this task remaining a low priority area [2]. Whenever a child is missing, parents report the case to the police. Once the missing child is reported, police officials get the photos of the missing child. A manual search will be conducted by police officials near the location of the disappearance or an announcement about the missing child will be made to the public. However, since most public places are equipped with cameras, the chances of a child kidnapping being captured by these cameras are high. As a result, this paper proposes an exclusive dataset for missing children, Smart Children Dataset (SCD), and a framework for detecting and retrieving abducted children based on a deep learning architecture known as YOLOv5 ChildLocate System derived from the baselines of You Only Look Once version 5. It reduces both the time and complexity of individually searching through huge amounts of CCTV footage.

To design or build a robust child-adult detection and retrieval framework, a thorough analysis of the current technologies for detecting a missing child is required. Currently, missing child detection technology relies on manual search processes [3]. There is no automated video surveillance-based detection framework available. To design a YOLOv5 ChildLocate System, one of the key aspects to analyse is the pre-developed object detection and retrieval frameworks that suit our field of research. Nevertheless, the existing object detection models identify a child and an adult as a single class of human beings. An approach like this is not suitable for missing child detection since it would increase the time complexity of the framework. In addition, a quick search algorithm is needed in order to achieve the performance required in our case-sensitive framework, as time complexity is an imperative factor. Persons are classified into three different categories, including a GirlChild, a BoyChild, and those older than 12 as an Adult. The initial stage of any object detection model is building the model using available benchmark datasets. Once trained, it is tested and validated for its performance based on the number of classes the model detects. To train and validate our ChildLocate System framework, we do not have any benchmark dataset except for a few web-crawled images available in Kaggle. Hence, we started collecting footage from various public places and web crawled images as part of our research. The dataset is called the Smart Children Dataset (SCD). Next, network architecture is required to train the model. There has been a remarkable expansion of both the theoretical foundations and practical uses of artificial intelligence in recent years [4]. Numerous studies have shown that object detection can be used successfully in a variety of fields like crowd estimation [5], rod safety [6], and pedestrian detection [7] etc.

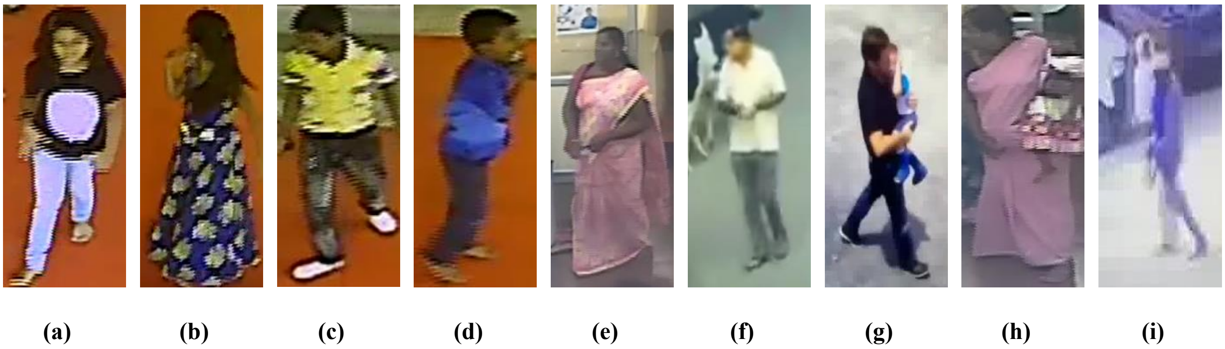


Figure 1. (a) To (i) represents the three different classes used for training the ChildLocate System framework. (a) & (b) represents the class: GirlChild, (c) & (d) represents the class: BoyChild, (e) & (f) represents the class: adult and (g), (h) and (i) represents the most challenging detection where the child is carried by adult (occlusion).

In the field of visual recognition, the development of Convolutional Neural Networks (CNN) [8] along with large-scale datasets has both made significant contributions. Although CNN-based models have outperformed humans on some benchmarks [9], deep learning techniques have not been widely used in fields where human lives are at stake, such as medicine or other case sensitive issues like crime detection or missing person detection. An improved YOLOv5 framework called YOLOv5 ChildLocate system is proposed for missing child detection and retrieval. The framework improves significantly compared to existing YOLOv5 [10] retrieval modules, which include models: Faster R-CNN, R-FCN, Mask RCNN, SSD, YOLOv3, and YOLOv5.

This paper contributes the following contributions:

- An exclusive child-centric dataset called Smart Children Dataset (SCD) collected from raw CCTV footages and web-crawled images which consists of about 7572 manually annotated child images and 3980 adult images.
- YOLOv5 ChildLocate system (Child-Adult detection and retrieval modules) is trained using three different classes: GirlChild, BoyChild and Adult. This helps the future retrieval process to address only the class specific to the victim reported which helps to reduce the time-delay in processing the entire footage with all classes.
- In existing object detection algorithms, a child carried by an adult is one of the challenges unaddressed. We trained the child-adult detection network with about 1780 images specific to this challenge.
- A novel child-adult detection model is developed from the baseline of the state-of-the-art algorithm YOLOv5 with lightweight BottleNeckCross Stage Polynomial feature extraction module with attention mechanism, and an additional prediction head.
- A multi-scale stacked convolution child-adult retrieval model with full image and part-based localization model using a four-channel input image.

This paper presents the ChildLocate System, a method for detecting and retrieving missing children, through extensive experiments on the SCD datasets. It discusses related works, implementation, evaluation, and validation of the system. The paper also includes a brief overview of related technologies and outlines future research directions. The study concludes with a discussion of related works and prevailing technologies for missing children detection and retrieval.

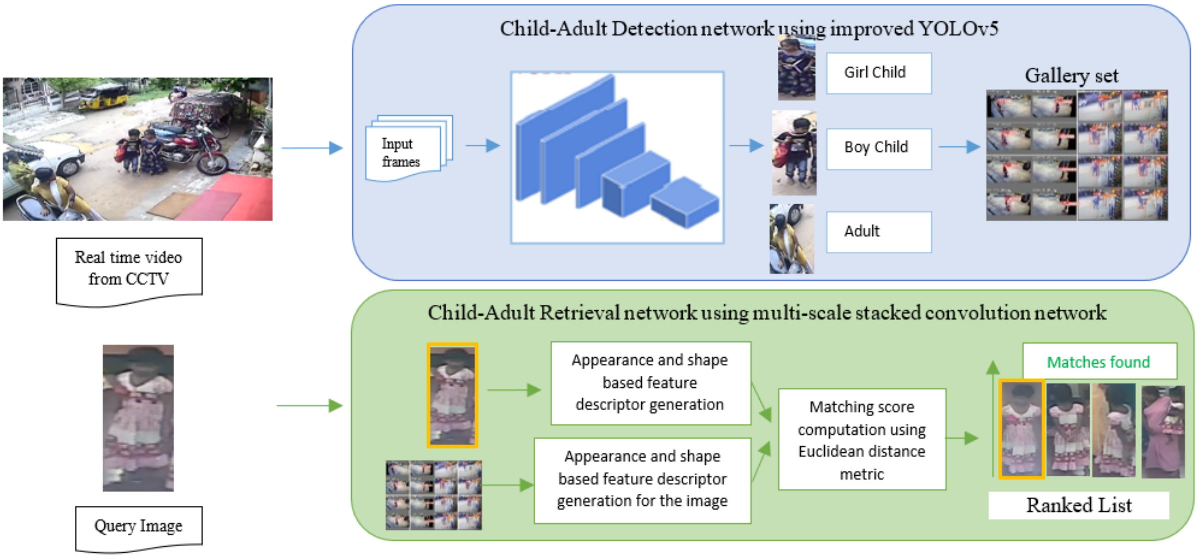

Real time video from CCTV
Child-Adult Detection network using improved YOLOv5
Input frames
Girl Child
Boy Child
Adult
Gallery set
Query Image
Child-Adult Retrieval network using multi-scale stacked convolution network
Appearance and shape based feature descriptor generation
Appearance and shape based feature descriptor generation for the image
Matching score computation using Euclidean distance metric
Matches found
Ranked List

Figure 2. The overall pipeline of Child Locate system. First, the Smart children dataset is used to train the improved child-adult detection model from the baselines YOLOv5 with lightweight Bottel neck CSP as backbone with attention mechanism, BIFPN as the neck and four-stage prediction head. Second, based on the training, the child-adult detector network, identifies the individual as one of the specific class, Girl Child, Boy Child or Adult. The detected results are stored in a gallery. Third, the query image of the missing child is given as input to the child-adult retrieval network developed from the baselines of MSCAN and MGCAM. Finally, based on the feature representation learning, the matching scores are generated and a rank list gives the best match from the gallery set.

RELATED WORK:

In response to the rise of deep learning, several benchmark datasets have been developed for the purpose of training and evaluating high-precision methods for finding people. It is well known that two-stage object detectors are highly accurate and capable of localizing objects very accurately. Faster R-CNN is the most representative of them all. But, they compensate for time as they process the image through two stages, where they determine the Region of Interest (RoI) in the first stage by dividing the image into smaller regions. Followed by bounding box regression on each of the region proposals. However, with the recent proliferation of high-speed GPUs and the widespread availability of image data, training a deep learning model [11] is more enriched, resulting in the introduction of many progressive one-stage CNN-based algorithms [12, 13] to improve object detection performance. The one-stage object detectors directly apply bounding box regression to the input image without any region proposals. Thus, one-stage object detectors are time efficient and work with real-time devices. However, the current object detection frameworks [14] focused on person detection, detecting both children and adults as one class (person) and did not have any datasets pertaining to children.

Two-stage object detectors:

Author	**Method**	**Backbone**	**Advantages**	**Limitations**	**Dataset**	**Person detection**	**Child-Adult detection**	**mAP (%)**
Girshick et al. 2014 [15]	RCNN	AlexNet	• First CNN based Object detector with region proposals Improves the quality of candidate Bounding Box Extracts high-level features	Complex multi-stage training 50 sec for an image Large disk storage and takes days to train even small dataset Geometric distortions	PASCAL VOC 2012	Y	N	53.3
He et al. 2015 [16]	SPP-net	ZF-5, Convnet-5, Overfeat - 5/7	It uses Spatial Pyramid Pooling layer Independent of input image's size/ aspect ratio Reduced computations	Fine-tuned only FC layer Similar complex multi-stage training as RCNN Longer training time	PASCAL VOC 2007, CalTech101	Y	N	59.2
Girshick et al. 2015 [17]	Fast RCNN	VGG16	An end-to-end training for classification and regression process. It shares the computations of CNN across all region proposals. A multi-task loss	Time consuming as it is still based on the traditional region proposal method taking 2 seconds per image	PASCAL VOC 2012	Y	N	66.0

Ren et al. 2016 [18]	Faster RCNN	VGG16 + ZF Net	Fast RCNN + RPN Detection accuracy increased by 3% Reduced test time: 5 fps	Tends to generate object like regions that includes background. Fails to detect multi-scale or multi-shape instances.	PASCAL VOC 2007 & 2012	Y	N	66.9
Dai et al. 2016 [19]	R-FCN	ResNet101	Fully Convolutional Neural network. Test speed of 170ms per image		PASCAL VOC 2007, MS COCO			83.6, 53.2 (AP@ 0.5)
He at al. 2018 [20]	Mask R-CNN	ResNet50 + FPN	Simple to train and flexible in posture estimation Misalignments overcome by RoI Align layer.	It fails to process temporal information of the images Detection accuracy is too low for low resolution inputs.	MS COCO	Y	N	70.4
Long et al. 2017 [21]	FPN	AlexNet, VGG net, and GoogLeNet	Detects small objects using multi-level feature maps with lateral connections.	It incurs false positive cases.	PASCAL VOC, NYUDv2, and SIFT Flow	Y	N	mean IoU : 62.2, 3.0 and 39.7

Inference: It is typical for traditional object detectors and region proposal-based frameworks to consist of several correlated stages, including generating region proposals, extracting features with CNNs, classifying, and predicting bounding boxes. These stages are typically trained separately, increasing the network's complexity. Here, a network's performance is completely dependent on the algorithm used to select candidate regions. In most cases, except for FCN, scale-invariance or local invariance are not addressed. There is still a lot of room for improvement in the training process and in the computation process. It is difficult to operate this pipeline in a real-time environment due to the fact that it is multi-stage and requires a large amount of memory.

One stage objects detectors:

Author	**Method**	**Backbone**	**FPS**	**Pros**	**Cons**	**Dataset**	**Person detection**	**Child-Adult detection**	**mAP (%)**
Redmon et al. 2016 [22]	YOLO	GoogleNet	45	Processes the entire image in a single stage. 45fps, suitable for real-time applications.	YOLO incurs localization error as the ratio of bounding box is merely learnt from the data Limitation in detecting more than one object per cell	PASCAL VOC	Y	N	63.4
Redmon et al. 2017 [23]	YOLOv2	DarkNet-19	67	Batch normalization layer and activation function helps with the convergence and regularization of the network. 40fps with reduced localization error	It does not show improvements in multi-label classification for more complex datasets containing many overlapping labels	PASCAL VOC	Y	N	78.6
Liu et al. 2016 [24]	SSD	VGG-16	59	Pre-defined set of anchor boxes at different scales and aspect ratio to detect objects at various scales and aspect ratio.	It has difficulty in detecting small objects.	PASCAL VOC	Y	N	74.3
Lin et al 2017 [25]	RetinaNet	ResNet + FPN	90	Overcomes the imbalance between positive and negative classes by using focal loss.	Detection speed is average	MS COCO	Y	N	61.1
Redmon et al. 2018 [26]	YOLOv3	Darknet-53	95.2	Better handles datasets with several overlapping labels that are increasingly complex. It uses multi-label classification that is independent logistic classifiers.	YOLOv3 performs better only with smaller object regions rather than larger regions. No significant change in accuracy than its predecessors.	MS COCO	Y	N	57.8

Bochkovskiy et al. 2020 [27]	YOLOv4	CSPDarknet 53	50	Trained using a single GPU. Bag of specials increases the accuracy of the model whereas Bag of Freebies modifies the training strategy.	Training time of the network is more even for a small dataset. The model's storage is about 245 MB The inference time is 22ms per image.	MS COCO	Y	N	65.7
Ultralytics 2021 [10]	YOLOv5	CSPDarknet 53 with Focus	140	Natively implemented in PyTorch, Inference time is 7ms per image. Reliable for exporting the net to any deployment environment Introduced a new data augmentation technique Automated the learning process of Bounding box.	Possible miss rates in multi-scale object detections. Small instance of object are considered as background.	MS COCO	Y	N	68.9

Inference: One-stage object detectors have reduced the time required for the object detection task which in turn reduces the complexity of training the network. A few network architectures respond well to multi-scale objects, but fail to detect small instances of the target object. Compared to other one-stage object detectors YOLOv5 is efficient for real-time deployment due to their upgraded PyTorch implementation by Ultralytics. They reduce training time and complexity, making them suitable for multi-scale objects. YOLOv5 also integrates anchor box selection, learning the most suitable anchor boxes for the dataset and using them during training.

Improved YOLOv5 models:

Author/ year of publication	**YOLOv5 variant**	**Model**	**Modifications**	**Objects detected**	**Person detection**	**Child-adult detection**	**mAP**	**Dataset**	**Optimizer**
Wang et al. 2022 [28]	YOLOv5s	YOLO-add	Modified Path Aggregation Network (PANet) to detect small objects combining FPN. Additional anchor boxes. The output of backbone is up sampled once more to generate a new feature map. A new detection layer	Detection of abnormal fish Behavior.	N	N	79.4	Collected about 564 images from aquariums.	SGD
Zhu et al. 2021 [29]	YOLOv5x	TPH-YOLOv5	Additional prediction head to address scale variance. Transformer prediction heads to detect overlap regions. Attention mechanism to RoI that covers more area	Drone captured images	Y	N	39.18	VisDrone2021	Adam
Yan et al. 2021 [30]	YOLOv5s	Improved YOLOv5	A modified BottleneckCSP module. SE attention module to improve feature extraction process The sizes of initial anchor boxes were modified.	Real-Time Apple Targets Detection	N	N	86.75	Image of apples from a specific species collected from a modern orchards.	SGD
Li et al. 2021 [31]	YOLOv5m	YOLO-FIRI	SK attention module inserted in CSP to enhance feature extraction. Four detection heads to address miss-rates in long range shooting	Infrared objects detection	Y	N	98.3	KAIST and FLIR	Not mentioned
Liu et al. 2022 [32]	YOLOv5l	YOLOv5-Tassel	Attention mechanism called SimAM is used before each head. Bi direction FPN is used to	Tassel detection in maize using RGB UAV Imagery	N	N	44.7	VisDrone2021	Adam

			fuse multi-scale features. Additional prediction added to detect small objects.						
Qi et al. 2022 [33]	YOLOv5m	SE-YOLOv5	A squeeze-and-excitation (SE) module is used in feature extraction process	To detect the tomato virus diseases	N	N	91.07	Collected images of tomato plants using mobile phones.	Not mentioned
Sun et al. 2022 [34]	YOLOv5s	SE-YOLOv5	Backbone's feature extraction is enhanced using a SE attention module. A varifocal loss function is used to address dense predictions. New data augmentation techniques introduced.	To detect the quality of the boiler plant's inner wall.	N	N	84.3	176 images collected from power boiler plant's inner walls.	Adam

Inference:

It is evident from the survey that there isn't any child-centric dataset or a deep learning based algorithm to detect the missing children. The existing algorithms consider both child and adult as a single class. Such scenarios can raise the network's temporal complexity. YOLOv5s has the most basic structure and runs quickly, meeting the need of real-time detection. When the targets are tiny or obscured, the detection precision of YOLOv5s diminishes. By fusing higher-level features with various attention mechanisms, this problem can be resolved. Hence, we propose a novel child-centric dataset and a deep learning based framework from the baselines of YOLOv5 to detect missing children using the given query image, addressing one of the challenges of occlusion where a child carried by an adult. It integrates a lightweight BottleNeckCSP module with CBAM, an improved feature aggregation module with a BiFPN. Moreover, an additional prediction head is used to detect children who appear generally smaller than adults

Person retrieval models

The goal of person retrieval model is to provide a ranked list of gallery photos of people that are closest comparable to a given query image of a subject of interest. If there exist no similarity between the give query image and the gallery set comprising on a whole all the instances classified images, then, it is concluded that the appearance of the query person is not present in corresponding CCTV footage. The primary goals of any person retrieval techniques are (1) the creation of a robust feature learning process for image representation, and (2) an efficient distance metric for matching identical individuals in close proximity to one another and distinguishing between them and other individuals at a further distance. The recent person retrieval methods [35] have attained by deep learning algorithms. Previous deep learning algorithms learn features for each individual in a pair of images using convolutional features or Fully Connected (FC) features [36] before applying a deep matching function. Whereas few works immediately learn image representation, and compares these results with the Euclidean metric. But, nowadays, methods that rely on deep learning for person retrieval often train both a representation of the individual and a similarity (distance) metric simultaneously. These strategies typically train a deep classification network to learn the ID-discriminative Embedding (IDE) feature [37]. The feature representation learning is of two types, the global feature representation learning and local feature representation learning. The global feature representation learning extracts a feature vector for each person image, whereas local feature representation learning extracts part-aggregated features. Even after years of work, the existing algorithms face difficulties in addressing the wide range of pose variations, background clutters and illumination effects. Hence, to enhance the feature representation learning process, several improvements are made that combine both global and local feature representations. They include methods that increase the richness of the training data, where the adversely occluded samples are generated for augmentation. Other publications have attempted to use multi-scale context or multi-resolution technique for retrieval. The aforementioned techniques rely on the entire image as input, which might still be negatively impacted by factors such as background clutters and posture variances. As a means of more thoroughly learning pedestrian representations, a number of approaches based on different body regions or parts have been suggested. Several posture based algorithms [38, 39, 40] have been presented in response to recent advances in pose estimation. These techniques show significant results, and they have demonstrated that eliminating backgrounds assists in improving the retrieval performance. Instead of using rigid body-parts, using deformable or latent body parts improved the elimination of background clutters. We propose a binary segmentation based methodology of the full body image, in addition to deformable body parts based background clutter elimination to further improve the global and local feature representation learning process. The background features are greatly suppressed by this approach.

METHODOLOGY:

YOLOv5 network architecture:

YOLOv3 and YOLOv4 are one-stage object detectors that follow the working principle of the heuristic You Only Look Once object detector. Each grid cell in the input image is responsible for predicting the confidence score as well as locating the target's bounding box. The conditional class probability is given in **Eq. (1).**

$$P_c(class_k|object) * P_c(object) * IoU_{B_{pb}}^{B_{gt}} = P_c(class_k) * IoU_{B_{pb}}^{B_{gt}} \qquad (1)$$

Where k represents the number of classes, B_{gt} represents the ground-truth bounding box area and B_{pd} represents the predicted bounding box area of the detected object that are used to calculate the Intersection of Union (IoU). When the grid contains the object, conditional class probability is 1, otherwise, 0. They follow the working principle of the heuristic You Only Look Once object detector. They use non-maximum suppression (nms) to enhance the predictions generated.

YOLOv5 is a novel network model that improves the performance of its predecessor YOLOv4. It is available in four variants, including small, medium, large, and extra-large. The CSPDarknet is the backbone, while PANet is the neck. The CSPDarknet addresses the problem of repeating gradient flow by updating gradient changes into the feature map. Focus enhances the receptive field, BottleneckCSP extracts feature information, SPP separates important contextual features, and PANet combines high- and low-level features for a more robust feature map. YOLOv5 outperforms YOLOv5x, YOLOv5m, and YOLOv5l in terms of detection accuracy and computational complexity. However, the detection precision decreases when targets are small or occluded. To address this, a lightweight CSPBottleNeck module and improved feature fusion based on BiFPN are added. The YOLOv5 ChildLocate system detects occluded child and adult features quickly, suitable for integrating in CCTV cameras that don't support higher GPUs.

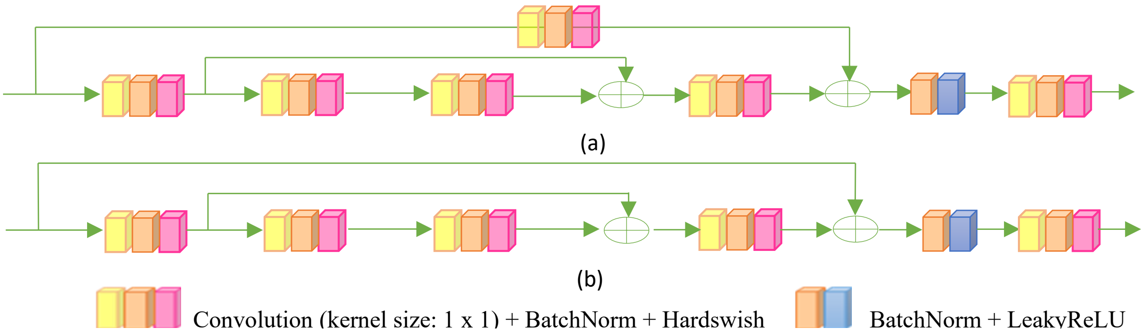


Figure 3. (a) BottleNeckCSP module of state-of-the-art YOLOv5, (b) Lightweight BottleneckCSP of ChildLocate

Improvement of the YOLOv5 network architecture:
One of the issues in using the dense and residual blocks is redundant gradients. CSPNet helps tackling this problem by truncating the gradient flow. First, CSPDarknet53 is deployed as the backbone of YOLov5 to extract the child-adult feature maps. The features maps of the base layer are divided into two by the CSPNet strategy and a cross stage hierarchy merges them together. This increases the inference speed of the network, as it reduces the number of parameters which in turn reduces the amount of FLOPs i.e. computations required. High Inference speed is one of the vital constraints in the real-time object detection models especially in such case-sensitive retrieval model, as they require a quick search framework to find the missing child using YOLOv5 ChildLocate System. YOLOv5 scales the original image to 1280 x 1280 pixels as input. For different feature maps, the backbone uses the BiFPN instead of the original PANet. It's followed by the attention module, CBAM. The scale-variance at different zones are addressed with an additional detection head included in the predictor in order to increase its multi-scale detection accuracy using NMS post-processing. These additional feature maps, addresses the problem of missed and false detections resulting from data captured from a distance. As part of its prediction process, YOLOv5 uses anchor boxes over the feature maps. These anchor boxes predict the final confidence score, class, and bounding box of the confidence score derived from the feature maps. The detection accuracy, inference speed, and number of network parameters of the proposed network model show a qualitative enhancement over the existing object detection model.

Improvement of Backbone Network
Lightweight BottleneckCSP module:
A backbone module plays an important role in the feature extraction process in any object detection network. It is then, these features are transmitted to the neck and head of the network. Thus, the backbone is indispensable since it is responsible for a large portion of the network's computations. Lightweight BottleNeckCSP module optimises and improves the backbone component of YOLOv5 architecture. The object detector used in ChildLocate System must not only quickly detect the missing child in a number of scenarios and complex public environment, but also the model's size should be reduced as much as possible to facilitate its integration in hardware devices, particularly CCTV.

The BottleneckCSP module consists of about six convolutional layers, two exclusively in the BottleNeck module and remaining four in the CSP module. Although the convolutional layer is one of the main building block that transforms the input to extract the features from it, the kernel/filter sums up a huge number of parameters in the network. In this research, an enhanced lightweight structure of the BottleneckCSP module is implemented. After removing the convolutional layer present in the BottleneckCSP module's cross stage, the input and output feature maps are concatenated directly in depth. This improved BottleneckCSP module helps in reducing the number of parameters to a greater extends. And to compensate for BottleneckCSP's lightweight characteristics, which may impede deep image feature extraction; four BottleneckCSP modules are substituted for their coterminous counterparts in the original backbone network. Thus, ChildLocate System is provided with a lightweight and enhanced network design to achieve high accuracy in target detection.

Convolutional block attention module (CBAM).

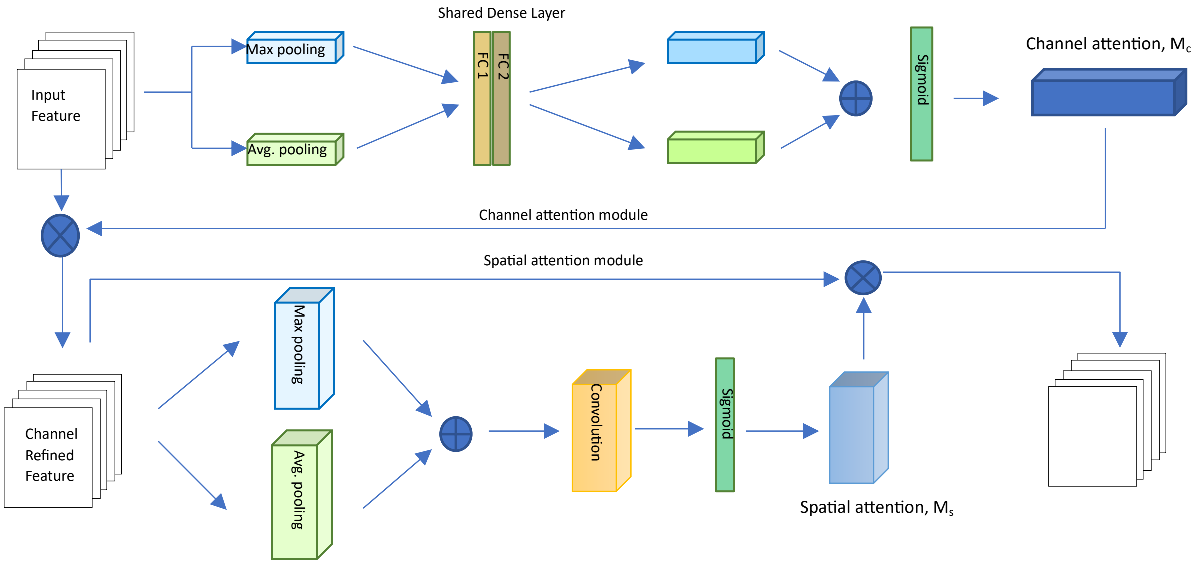


Fig. 4. Convolutional Block Attention Module (CBAM)

The attention module called the Convolutional block attention module is integrated in the network that increases detection accuracy of the child-adult detection model. Thus, it differentiates the target from the background and better extracts the properties of a child-adult. It is a simple lightweight and efficient module that is trained with most CNN architectures in an end-to-end manner. To attain an adaptive feature refinement, it multiples the attention map from both the spatial and channel dimensions with the input feature map. The model's efficacy is significantly improved after integrating CBAM as it enhances beneficial features and suppress unimportant ones. Since this module's computation is modest, it is included in the backbone network of the modified YOLOv5s architecture in this study to improve the model's detection accuracy.

Improved Feature Fusion based on BiFPN:

The detection head receives the output feature map from the neck, which consists of a sequence of feature aggregation layers comprising mixed and merged image features that are mostly used to produce Feature Pyramid Network (FPN) [41]. As a result of this the feature extractor integrates a novel FPN structure that improves the bottom-up route enhancing the transmission of low-level features and the detecting objects at different scales. Children’s body built and height varies when they cross each zone of the field of the camera. Spatially the area occupied by children are lesser than adults, they become even smaller when they are in the farthest/ later zone compared to the middle and early zones (**see Fig. 4**). The feature pyramid network has been found to be an effective technique for detecting objects at different scales. A bottom-up pathway, a top-down pathway, and lateral connections constitute this network. The information flow between the down and up layers can be shortened by using an additional information flow introduced to facilitate the fusion of multi-scale feature maps. Despite this, the contributions from different layers at different resolutions to a fused feature are likely to be unequal. In order to overcome this problem, a network that allows rapid multi-scale feature fusion can be used. This is when BiFPN was introduced. As a result of the advantages of BiFPN, we propose to incorporate it into YOLOv5. This will increase the ability of feature fusion in the four detection heads presented in this article. Bi-FPN introduces learnable weights, enabling the network to learn the importance of different input features, and repeatedly applies top-down and bottom-up multi-scale feature fusion. Compared with Yolov5′s neck PANet, Bi-FPN has better performances with less parameters and FLOPS. Furthermore, different feature fusion strategies bring different semantic information, resulting in multi-scale detection results. Therefore, the same target may be precisely detected regardless of its size or scale when crossing each zone of the shooting area covered by the camera.

Additional Head addressing Scale-Invariance:

The Smart-Children-Dataset consists of many extremely small instances of children present in the far zone, thus requiring an additional prediction head for tiny object detection. Combined with the other three prediction heads, our four-head structure can ease the negative influence caused by violent object scale variance. The prediction head (head No.1) added is generated from low-level, high-resolution feature map, which is more sensitive to tiny objects. After adding an additional detection head, the performance of tiny objects detection significantly improves. As a result, levels 2–5 comprise the input features. Thus, transformation to aggregate four different features is designed to output four-scale fused features, which is given by eq. 2.

$$\begin{aligned}
P_5^{out} &= Conv(P_5^{in}) \\
P_4^{out} &= Conv\ (P_4^{in} + Resize\ (P_5^{out})) \\
P_3^{out} &= Conv\ (P_3^{in} + Resize\ (P_4^{out})) \\
P_2^{out} &= Conv\ (P_2^{in} + Resize\ (P_3^{out})) \qquad (2)
\end{aligned}$$

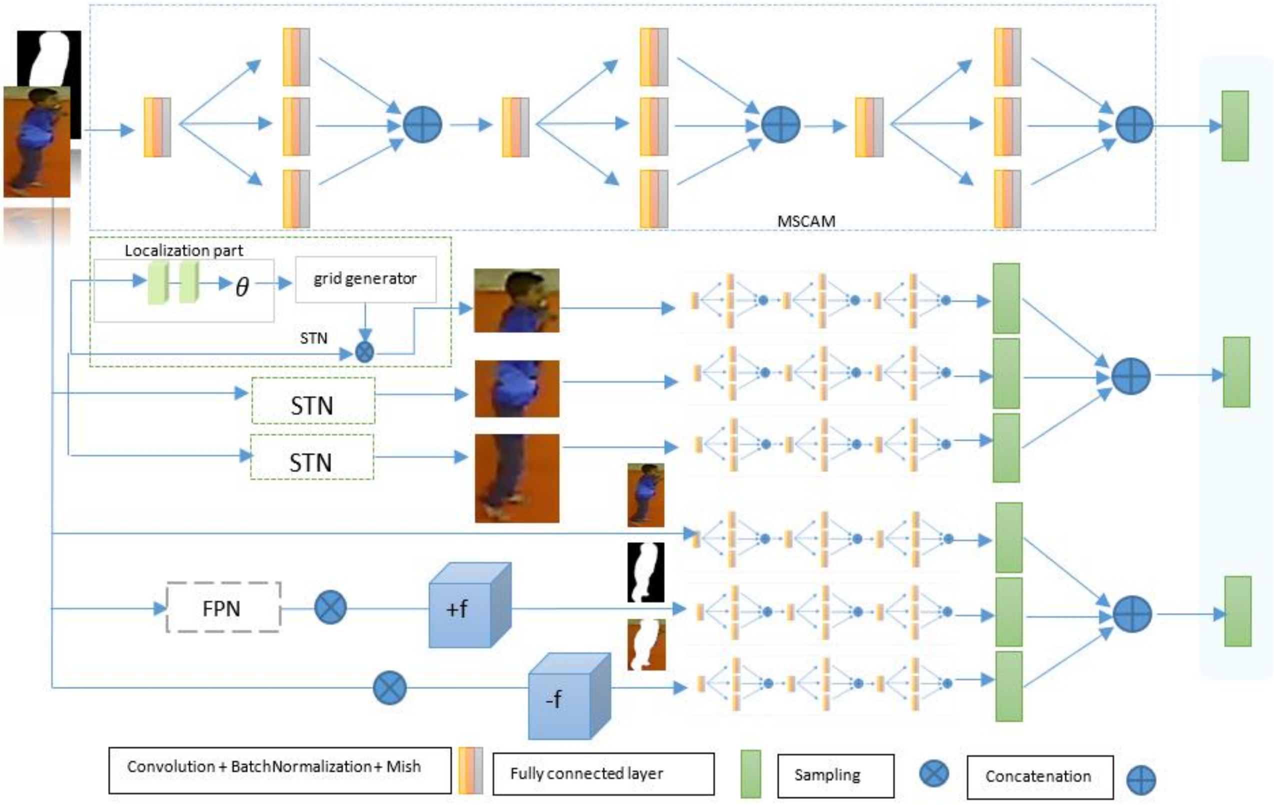


Fig. 5. The overall child-adult retrieval framework

For the purpose of matching four different scale feature maps, Resize represents an up-sampling and down-sampling operation, while Conv represents a convolutional operation to extract features.

As BiFPN combines bidirectional cross-scale connection and fast normalized fusion, the feature map output for each layer is given by eq. 3.

$$P_i^{td} = \mathrm{Conv}\left(\frac{w_1 \,.\, P_i^{in} + w_2 \,.\mathrm{Resize}\,(P_{i+1}^{in})}{w_1 + w'_2 + \epsilon}\right)$$
$$P_i^{td} = \mathrm{Conv}\left(\frac{w'_1 \,.\, P_i^{in} + w'_2 \,.P_i^{td} + w'_3 .\mathrm{Resize}\,(P_{i-1}^{in})}{w_1 + w'_2 + w'_3 + \epsilon}\right) \qquad (3)$$

Where $i = 2{,}3{,}4$ and P_i^{td} are the intermediate features at level i, respectively, on the top-down pathway. The output feature of P_5^{out} can be described as follows:

$$P_5^{out} = \mathrm{Conv}\left(\frac{w'_1}{\cdot} P_5^{in} + w'_2 \cdot \mathrm{Resize}(P_4^{out}) w'_1 + w'_2 + \epsilon\right) \qquad (4)$$

Based on these operations, the output feature map integrates the input feature map and the intermediate feature map with different scaling features. Thus, it further improves the feature fusion process after introducing four-detection heads.

Table 3.
Performance analysis (LAMR) of YOLOv5 ChildLocate system under various illumination condition of Child-Adult Class

Training set	Test set			
	Outdoor	Indoor	Day	Night
Outdoor	7.2	17.3	25.8	36.2
Indoor	9.5	38.2	14.6	19.3
Day-Night	12.9	22.6	17.7	27.9
Outdoor-Indoor	26.7	18.3	22.8	33.4

Retrievalframework:

Image retrieval is the process of matching a given query image against the database of images. Here, in our model, the photo of the child went missing is given to the police officials. Based on this photo-query the retrieval system identifies the correct match from the database consisting of girl-child, boy-child and adult detected by improved YOLOv5 model. Since, we build a quick search retrieval system as we have trained the person detection model using three different classes, i.e. a gender-based child detection, girl-child, boy-child and adults. This reduces the search/retrieval time as a class-specific search will be carried out based on the query given as input as shown in the figure 1.

Four channel input image:

In image retrieval process, one of the most challenging task is to avoid the background features that affect the retrieval accuracy. This is overcome by using a four-channel image which consists of Red, Green, Blue and the binary mask of the image as the fourth channel based on the works of Chunfeng Song et al. [42]. It is the RGB channels which contain the appearance-based features and the binary mask of the original image which contains the shape-based features. This helps to distinguish the foreground region of interest (query) from the background based on the appearance-feature of the missing child. Further, since ChildLocate System will be integrated in highly complex environments which includes public places like airports, railway stations, bus stands, theatres and malls, the illumination effects differs from place to place. As a binary image operates at the pixel-level it tends to be more robust in addressing these complex environments. And also there are more possibilities for the change in cloth colour, or cloth-type of the missing-child. The only change invariant robust feature in such scenarios is the gait of the missing-child. Thus, the four-channel image provides the body-shape information which is an important gait-feature and thus helps to extract both low-level and high level features that helps to identify a missing child. This binary mask of the RGB image is obtained using FPN.

Stacked Dilated Convolution Network:

The four-channel image is given as input to a multi scale dilated convolution-based network [43] which helps to extract the receptive fields in the dilation ratio of 1:2:3 to obtain the contextual features at various scales. Dilation ratios are changed from the usual kernel as they produce more redundant information, but whereas a dilated convolution produces a larger receptive field with redundant pixel only at the centre. Each convolution operation is followed by a BatchNormalization and a mish activation function. The network with Mish improved Top-1 test accuracy by 1.385% over Swish and 0.8702% over ReLU. As a final output of the stacked convolution network, they are concatenated along the channel-axis.

Part-Localization Network:

As the spatial transformer networks (STN) [44] are fully differentiable and easily Integrated into any convolutional neural network, they are used for deformable part-based learning and localization. Instead of using rigid-body parts of the person as it greatly affects the pose variations of a person, deformable part-based learning is adopted. STN operates in two-stages, the transformation parameters are learnt through spatial localization followed by image sampling using interpolation kernel. We use the bilinear-interpolation kernel of the STN where the four transformation parameters representing scaling and translation that helps to crop the full image of the person. Each body-part is learnt separately by a STN with three constrains on the transformation parameters for proper convergence of the network. The three constraints are the centre-region constraints (L_{cen}), the value range constraint (L_{pos}), and the inner region constraint (L_{in}) as shown in equation (5). The loss for the transformation parameters is given by equation ( ).

$$L_{cen} = \frac{1}{2}\max\{0, (t_x - C_x)^2 + \left(t_y - C_y\right)^2 - \propto\}$$
$$L_{pos} = \max\{0, \beta - s_x\} + \max\left\{0, \beta - s_y\right\}$$
$$L_{in} = \frac{1}{2}\max\{0, \|s_x \pm t_x\|^2 - \gamma\} + \frac{1}{2}\max\{0, \left\|s_y \pm t_y\right\|^2 - \gamma\}$$
$$L_{loc} = L_{cen} + \varepsilon_1 L_{pos} + \varepsilon_2 L_{in} \tag{5}$$

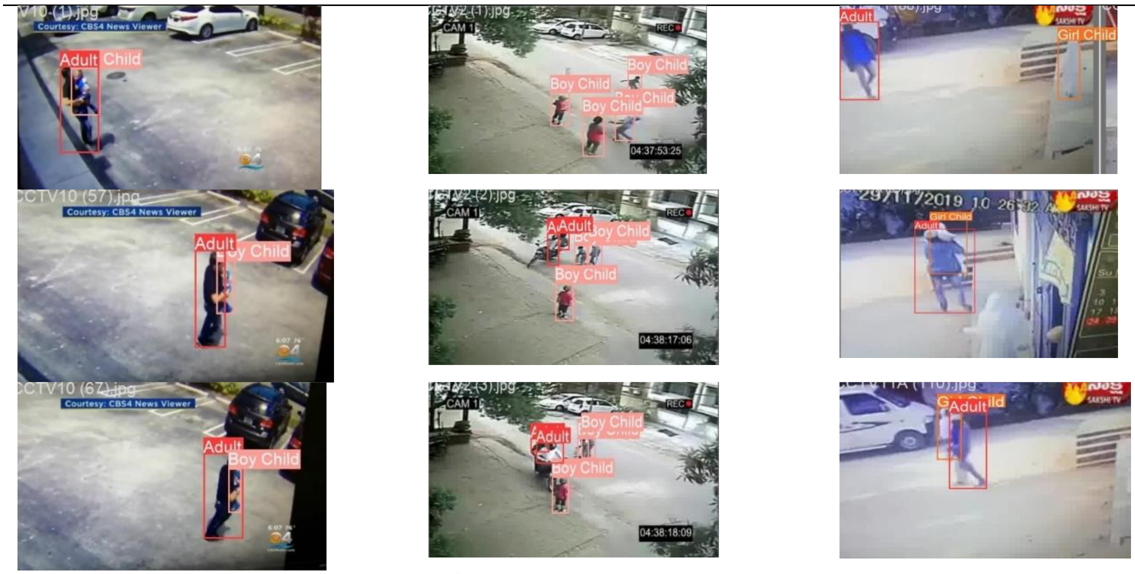
Where, t_x, t_y are the horizontal and vertical translation parameters, s_x, s_y are the horizontal and vertical scale transformation parameters, $C_x, C_y$are the prior center points of each body part. $\propto and\ \beta$are the threshold parameters, γ is the boundary parameter and $\varepsilon_1\ and\ \varepsilon_2$ are the hyperparameters.

Overall Missing Child Retrieval (MCR) network architecture:

The multi-scale mask guided convolution network works as three streams. In the first stream, the 4-channeled full-body image is given as input to the stacked multi-scale convolution layers which learns the global features of the Region of Interest followed by a fully connected layer with 128 dimensional feature vectors. Parallel, the second stream, STN is used to extract the part-based features maps from each body part separately by stacked dilated convolution followed by the FC layer. Finally, the third stream, FPN followed by a STN where body-part based binary mask is used to extract more complex features. The region level triplet loss (t_L) is used for the feature learning process from three-different streams as given in equation (6). We integrate the features maps from three streams for the representation learning processes of four channel input images with RGB channels and their corresponding binary mask into a joint structure for the person retrieval task. Using RGB channels with a binary mask enhances the feature learning process to a greater extend, even in worst case when the binary image generated is wrong, the multi-stack convolution blocks can still learn the features from RGB channels.

Table 4.
Qualitative validation of YOLOv5 ChildLocate system (Red: Adults, Orange: GirlChild, Blue: BoyChild)

True Positives



$$t_L = \|f_F - f_{Rp}\|_2^2 + \max\{\left(m - \|f_F - f_{Bp}\|_2^2\right), 0\} \qquad (6)$$

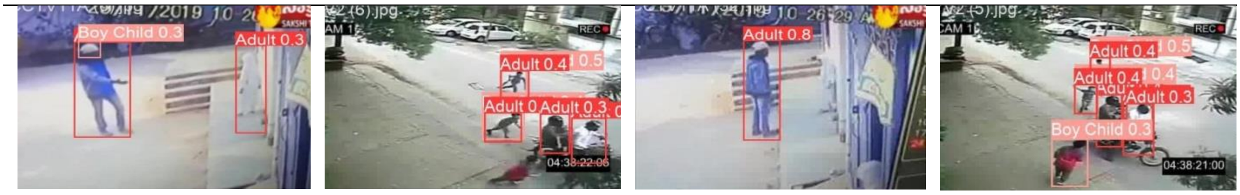
Where, f_F, f_{Rp}, f_{Bp} represent the 128 dimensional feature vectors of the three streams, m is the margin parameter set to 10. Finally, to measure the distance between the query and the match, L2 normalized person representation is used.

Training and Performance Evaluation:

This section evaluates the proposed method ChildLocate System qualitatively and quantitatively compared to the conventional methods. As far as from our knowledge there isn't any child-centric benchmark datasets available except for a few that are found in Kaggle. These few Kaggle datasets are limited to only about 500 web-crawled datasets. Hence, we introduce a new dataset called Smart children dataset (SCD) of children covered under surveillance (described under dataset subsection). Experimental results demonstrate that ChildLocate System framework achieves state-of-the-art performance on three datasets. In the next subsections, we first introduce the datasets and implementation details. And then we investigate the effects of each component of our improved YOLOv5 network presenting a series of ablation experiments to inspect the impact of components and several variants of our model

Table 5.
Qualitative validation of YOLOv5 ChildLocate system (Red: Adults, Orange: GirlChild, Blue: BoyChild)

False Positives



Dataset Acquisition and Augmentation:

Smart Children Dataset Acquisition

In this study images of children and adult from various public places like theatres, malls and school premises are collected and used for the research purpose. According to the literature survey, there isn't any dataset that exists for children. Hence, we built our own dataset called the Smart Children Dataset. We collected 10 hours of video sequence in total, for an average of 0.5 hours per day recorded from various exists of public places and other web-crawled data. The dataset also consists of video footages collected from real-world news reports. The dataset will be made publicly available in github. Adolescents (10 and 19 years of age according to World Health Organisation) and individuals above 20 years are considered as adults. The distribution of adults and children in the acquired dataset is about 60% and 40%. The common spacing between each individual is only about few feet, and the average height of children is between 68cms and 127cms. The dataset includes images under various illumination conditions. Various CCTV footage includes both day and evening time zones. The video footages collected from various CCTV cameras installed at different tilt angles and camera height. The region of coverage of the CCTV camera called the field of view is typically around 1.2 meters. The subjects covered include various pose variations: children carried by adults, children occluded by adults, children occluded by other objects in scene. The dataset also includes different light condition that cover indoor and outdoor environment both in natural and artificial lighting effects. Dome cameras and bullet cameras are widely used in the environment that includes Hikvision and dahua cameras.

SCD Description and File Format

Since the development of the deep-learning-based target detection model is achieved by the training of a large amount of image data, the 11,480 collected child-adult images required augmentation (section 3.1.3). (given in section 2.1.3). Out of a total of 11,480 photos, 3444 (30% of the training set) are chosen at random to serve as the test set, while the remaining 8037 are used as the training set. Second, we reduce the size of the original 8037 images in the training set by a factor of five in order to improve the training efficiency of the child-adult detection network. We use "MakeSense.ai" to manually annotate the images by adding rectangle boxes around the child and adult subjects in the compressed images. Images are categorised according to the smallest surrounding rectangle of each category, with the goal of including as little background as possible within the rectangle. The children in the image are separated into two classes—"Girl-Child" and "Boy-Child"—based on their gender, while adults are grouped together under the umbrella term "Adult." The .yaml format files are generated after the annotation are exported and saved. Finally, data augmentation is applied to the training set's images to enrich the data set, allowing for more accurate feature extraction for each person in the labelled categories and preventing the model from being over-fit to the training data.

Table 6
Performance comparison of YOLOv5 ChildLocate system with the state-of-the art One stage Object detectors on SCD

Object detection networks	Frames Per Second (FPS)	Size of Model (MB)	mAP (%)
ResNet50 [45]	0.5	23.48	72.91
Faster R-CNN[18]	5	41.20	75.32
R-FCN [19]	8	50.90	79.46
SSD [24]	9	56.80	82.72
YOLOv3 [26]	35	235.0	82.68
YOLOv5 [10]	48	14.0	83.86
YOLOv5 ChildLocate system	52	12.9	86.52

Albumentation
To increase the diversity of dataset and improve the generalization ability of the of the improved YOLOv5 network which in turn increases the accuracy and prevent the model from overfitting data augmentation techniques is used. A number of image augmentation techniques are exploited for the 8050 images of training set using Albumentation python library. Albumentation supports all computer vision tasks with a simple unified API that works with all data type. The library consists of about 70 types of augmentation techniques including both pixel level and spatial level transforms. Channel shifts –ShiftRGB (r/g/b_shift_limit = -20:20), ShiftHSV, RandomBrightness (limit = -0.2:+0.2), RandomContrast, RandomGamma, and Gaussian Noise (var_limit = 10:50, mean = 0). These are the pixel-level transforms of the SCD images. The spatial level transform of the image dataset include HorizontalShift, VerticalShift, Rotate, RandomCrop, and RandomScale, ElasticTransform(alpha = 1, sigma = alpha_affine = 50), and SmallestMaxSize(max_size = 512).

Network Training:
The training process of YOLOv5 ChildLocate system is carried in Windows 10 operating system of HP Pavilion 15 Notebook Personal Computer, Intel(R) Core(TM) i7-5500U CPU, 2.40GHz Microprocessor, system memory of 8GB and NVIDIA GeForce 840M Graphic Device. The architecture is built with the Python 3.7 as the scripting language using the Anaconda compiler. Before the model in this article starts to be trained, its hyper-parameters need to be initialized. It is trained using the input image of size 630x630x3 with the Stochastic Gradient Descent (SGD) optimizer and a batch size of 4 with 250 epochs. During the training process, the network continuously updates the parameters to accelerate network convergence and prevent overfitting. Testing of the improved YOLov5 model is carried out using the weight file saved from the training process to evaluate the model's performance. The final output is the localizing the detected classes (boy-child, girl-child and adult) with bounding boxes and the percentage of probability of the detected class.

Training Results
According to the training performance of the proposed model, the training-loss initially reduced rapidly 120 epochs and, in general, it be likely to remain constant after 230 epochs. As a result, after 320 epochs of training, the model output is determined as the child adult recognition model for the YOLOv5 ChildLocate architecture. Figure 11b also depicts the mAP (mean average precision) of the training and validation sets.

Performance metric
Child-Adult detections are presented by the following attributes: the precited class, the accompanying bounding box, and the confidence score, which is often a number ranging from 0 to 1 indicating how assertive the detector performs prediction. The evaluation is based on a set of ground-truth bounding boxes representing rectangular portions of an image containing objects of the class to be identified, as well as a set of detections predicted by a model. To analyse the performance of the ChildLocate System, evaluation metrics such as Log Average Mean Precision, Miss Rate, False Positive Per Image, Jaccard Index, Precision, Recall, and mean Average Precision are used, following the

Table 7.
Ablation study on YOLOv5 ChildLocate system

Method	#Param	mAP (%)
YOLOv5s (Baseline)	46.1M	84.8
+ Lightweight CSPNet	44.2M	85.2[+1.4]
+ BiFPN	50.6M	87.6[+2.3]
+ CBAM	51.9M	89.2[+2.1]
+ 4 detection heads	52.5	91.2[+2.3]

Table 8.
The performance analysis with various attention modules

Attention Module	#Param	mAP (%)
SimAM [46]	51.8M	32.7
Shuffle AM [47]	51.8M	27.5
SELayer [48]	51.9M	22.4
ECALayer [49]	51.8M	20.9
CBAM [50]	51.9M	42.3

Table 9.
Performance analysis of YOLOv5 ChildLocate system on various test cases.

Test set	**Number of Images**	**Precision** (%)	**Recall** (%)	**mAP** (%)	**F1** (%)
BoyChild	1490	86.62	93.22	88.12	88.61
GirlChild	1260	82.97	81.24	78.51	84.32
Adult	2320	93.65	98.43	95.98	96.92
BoyChild-Adult	750	83.29	84.62	87.32	86.87
GirlChild-Adult	620	78.25	80.34	71.77	82.87
Total	**6440**	**91.82**	**87.42**	**90.46**	**92.65**

widely acknowledged technique of ECP [51]. Using similar evaluation parameters, we evaluate and compare all approaches.
Log Average Miss Rate (LAMR)

The log-average miss rate is the standard metric in person detection (LAMR). LAMR necessitates calculating the Miss Rate and False Positive per Image (FPPI). Given a detection confidence score threshold, miss rate (MR) can be calculated by dividing the number of true positives (N_{tp}) by the number of ground truths (N_g) as MR = 1 - N_{tp}=N_g; and false positives per image (FPPI) can be calculated by dividing the number of images by the number of false positives. It can be visualised in log-space by adjusting the confidence level. Finally, the log-average MR is calculated by averaging miss rates under 11 FPPI equally spaced in [102:100]. A lower MR reflects a better performance.

$$mr(c) = \frac{f_n\ (c)}{t_p(c) + f_n(c)}$$

$$fppi(c) = \frac{f_p(c)}{t_p(c) + f_p(c)} \qquad (7)$$

where, True Positive (t_p) refers to the number of positive classes identified from the trained classes i.e. girl-child, boy-child and adult; False Positive (f_p) refers to the number negative claases classified as positive, where the backgroundclutter or other similar contextual feature in the image is detected as child or adult; False Negative (f_n) refers to the actual child or adult that are left unattended by the ChildLocate System.

Jaccard Index/ Intersection of Union (J/IoU)
If a detected bounding-box is a true positive (t_p) or a false positive (f_p), it is represented by a ground-truth bounding box (B_{gt}) and a predicted bounding box (B_{pd}), respectively. A perfect match is considered when the area and location of the projected and ground-truth bounding boxes are the same, regardless of confidence level.

$$\mathrm{J}(B_{gt}, B_{pd}) = \frac{(B_{pd} \cap B_{gt})}{(B_{pd} \cup B_{gt})} \tag{8}$$

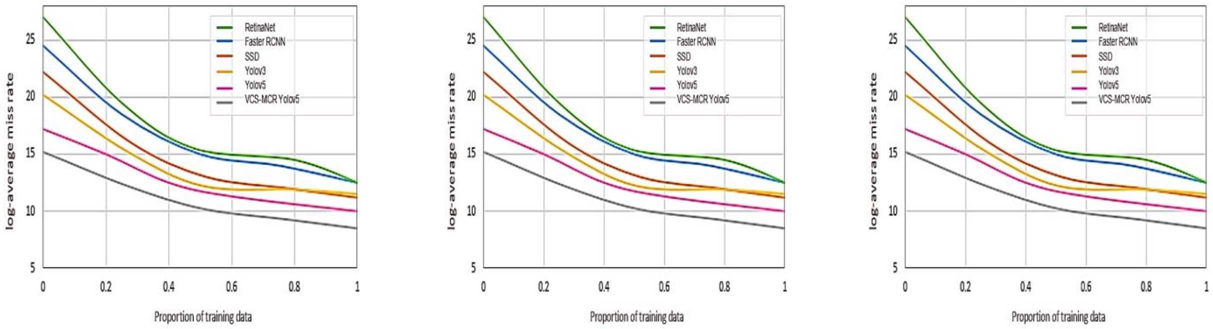


Fig. 8. Detection Performance of YOLOv5 ChildLocate System with different proportions of training data

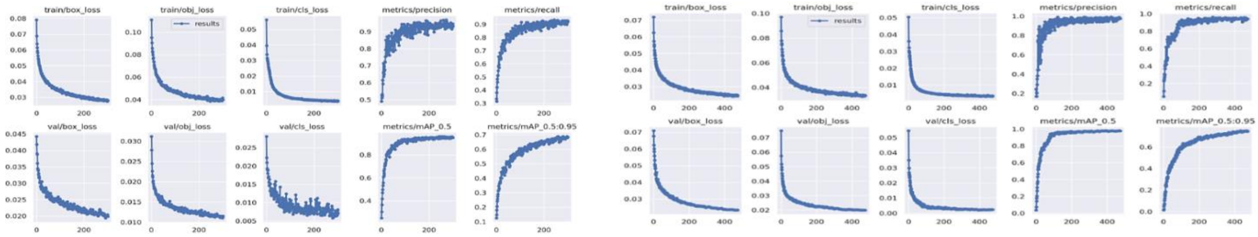


Fig. 9. Training and Validation results of the state-of-the-art YOLOv5 model and the proposed YOLOv5

RESULTS AND DISCUSSION:

The testing phase employs 2200 children images and 1300 adult images from the test set to validate the performance of the real-time child Locate system-based child-adult recognition model created in this study. There are a total of 5280 people in 3500 test set images, with 3168 of them being youngsters and 2112 being adults. Table 3 shows the specific recognition results of the improved YOLOv5 child-adult detector in the study, which show that the precision, recall, mAP value, and F1 score of the improved YOLOv5 are 86.62%, 93.22%, 88.12%, and 88.61%, respectively; for the GirlChild-class, the identification results are 82.97%, 84.62%, 78.51%, and 86.32%, respectively; for the adult-class, the identification The child-adult identification model's overall precision, recall, mAP, and F1 are 87.62%, 91.82%, 87.42%, and 90.46%, respectively. The improved YOLov5 network's performance for child-adult detection is evaluated. The model's mAP and LAMR are employed as assessment indicators. According to Table 4, the upgraded YOLOv5s child-adult detector suggested in the study has a higher mAP value, which is 1.9% higher than the original YOLOv5 network and 3.84% and 2.66% higher than SSD and YOLOv3, respectively. With an average detection speed of 52 frames per second, the improved YOLOv5s model is 1.42, 1.02, and 2.42 times faster than SSD, YOLOv3, and YOLOv5 networks. The retrieval network's performance is compared to that of six cutting-edge networks. According to the comparative results based on the ranking list and mAP, the proposed child-adult retrieval network outperforms the MSCAN by 0.32%. Three alternative distance metrics are employed to assess matching scores; the suggested child-adult retrieval's Euclidean distance measure exhibits satisfactory gains with 66.39% mAP. Consequently, the child-adult detection and retrieval results demonstrate considerable performance, indicating that the child-adult detection network meets the requirements of the real-time ChildLocate System.

Table 11.

Retrieval results of the ChildLocate System dataset

Methods	Rank1	Rank5	Rank20	mAP
PersonNet [52]	48.28	61.97	66.87	31.66
Spindle [53]	52.58	71.59	75.21	77.72
Re-Ranking [54]	60.51	73.18	67.22	43.50
DLPAR [55]	63.03	73.69	77.63	46.43
SVDNet [56]	65.84	75.93	79.24	48.63
GAN [57]	70.64	82.93	87.14	80.24
MSCAN [42]	66.88	80.03	84.25	86.24
MGCAM [43]	70.12	86.24	87.93	73.24
Ours	70.72	86.14	82.25	86.56

Table 12.

Retrieval results with three different distance metrics

Methods	Distance Metric	Rank1	mAP
MSCAN [42]	Euclidean	60.1	43.81
	XQDA	56.11	31.75
	Re-ranking	60.73	52.19
MGCAM [43]	Euclidean	61.83	42.66
	XQDA	64.52	42.32
	Re-ranking	65.72	61.29
Ours	Euclidean	70.66	66.39
	XQDA	70.34	51.02
	Re-ranking	72.06	66.27

Table 10.

Performance (mAP) of YOLOv5 ChildLocate system on different illumination conditions for various Test Case

	Test case					
	BoyChild		GirlChild		Adult	
Training	Outdoor	Indoor	Outdoor	Indoor	Outdoor	Indoor
GirlChild-Adult	78.3	87.5	80.1	83.9	89.2	90.1
BoyChild-Adult	86.5	88.2	75.2	84.5	91.2	92.3
Adult	90.3	92.1	88.5	90.5	93.9	95.6

CONCLUSION

We present missing child detection and retrieval framework that helps the law enforcement department to find a missing child based the query image collected. The proposed framework and the smart children dataset will be one of its first kinds in the literature of missing child detection systems as there doesn't exist any dataset or a robust algorithm based on quick search. The smart children dataset consists of about 10, 000 images of children and adult collected from various resources. The improved YOLOv5 model integrated in ChildLocate System performs well with is 86.4 (2.66) % increase in mean Average Precision with 52fps Compared to the YOLOv5 with 85.2% with 45fps. To validate the utility and functionality of the ChildLocate System, the proposed model is evaluated on the SmartChildren Dataset (SCD). Extensive experiments and ablation studies are performed (i) in various light conditions, (ii) with various camera sensors, and (iii) in various other use cases. The overall-performance of the proposed ChildLocate System shows 88.5%1 overall mean Average Precision. In future, the missing child detection framework based on other soft-biometric traits will be carried out to improve the retrieval accuracy of the system.

REFERENCES

1. [online] Available: https://www.nhrc.nic.in/Documents/Reports/misc_MCRReport.doc
2. [online] Available: https://rakshakfoundation.org/wp-content/uploads/formidable/FinalReport_ProjectID_122_Sameer_Shaikh.pdf
3. [online] Available: http://www.missingindiankids.com/missing/
4. Zhou, Quan, et al. "Knowledge implementation and transfer with an adaptive learning network for real-time power management of the plug-in hybrid vehicle." IEEE Transactions on Neural Networks and Learning Systems 32.12 (2021): 5298-5308.
5. Chen, Xinyue, et al. "Adversarial scale-adaptive neural network for crowd counting." Neurocomputing 450 (2021): 14-24.
6. Liu, Wei, et al. "Automated vehicle sideslip angle estimation considering signal measurement characteristic." IEEE Sensors Journal 21.19 (2021): 21675-21687.
7. Li, Guofa, Yifan Yang, and Xingda Qu. "Deep learning approaches on pedestrian detection in hazy weather." IEEE Transactions on Industrial Electronics 67.10 (2019): 8889-8899.
8. Yamashita, Rikiya, et al. "Convolutional neural networks: an overview and application in radiology." Insights into imaging 9.4 (2018): 611-629.
9. Kaiming, He, et al. "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification kaiming." Biochemical and Biophysical Research Communications 498.1 (2018): 254-261.
10. [online] Available: https://github.com/ultralytics/yolov5
11. Zaidi, Syed Sahil Abbas, et al. "A survey of modern deep learning based object detection models." Digital Signal Processing (2022): 103514.
12. Aziz, Lubna, et al. "Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review." IEEE Access 8 (2020): 170461-170495.
13. Ansari, Mohd, and Dushyant Kumar Singh. "Human detection techniques for real time surveillance: A comprehensive survey." Multimedia Tools and Applications 80.6 (2021): 8759-8808.

14. Jiao, Licheng, et al. "A survey of deep learning-based object detection." IEEE access 7 (2019): 128837-128868.
15. Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
16. He, Kaiming, et al. "Spatial pyramid pooling in deep convolutional networks for visual recognition." IEEE transactions on pattern analysis and machine intelligence 37.9 (2015): 1904-1916.
17. R. Girshick, Fast R-CNN, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp.1440–1448.
18. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, arXiv:1506 .01497, 2016.
19. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, arXiv:1703 .06870, 2018.
20. Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
21. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.779–788.
22. J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, https://arxiv.org /abs /1612 .08242v1.
23. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot MultiBox detector, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, 2016, pp.21–37.
24. J. Redmon, A. Farhadi, YOLOv3: an incremental improvement, arXiv:1804 .02767, 2018.
25. Bochkovskiy, Alexey, Chien Wang, and Hong Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection." *ArXiv*, (2020). Accessed January 24, 2024. /abs/2004.10934.
26. Wang, He, et al. "Real-time detection and tracking of fish abnormal behavior based on improved YOLOV5 and SiamRPN++." Computers and Electronics in Agriculture 192 (2022): 106512.
27. Zhu, Xingkui, et al. "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
28. S. Li, Y. Li, Y. Li, M. Li and X. Xu, "YOLO-FIRI: Improved YOLOv5 for Infrared Image Object Detection," in IEEE Access, vol. 9, pp. 141861-141875, 2021, doi: 10.1109/ACCESS.2021.3120870.