Deep Learning based Human Detection in Privacy-Preserved Surveillance Videos

Muhammad Jehanzaib Yousuf
Software Research Institute
TU Shannon: Midlands Midwest
Athlone, Ireland
a00278629@student.ait.ie

Nadia Kanwal
School of Computing and Mathematics
Keele University
United Kingdom
n.kanwal@keele.ac.uk

Mohammad Samar Ansari
Faculty of Science and Engineering
University of Chester
United Kingdom
m.ansari@chester.ac.uk

Mamoona Naveed Asghar
School of Computer Science
National University of Ireland Galway
Ireland
mamoona.asghar@nuigalway.ie

Brian Lee
Software Research Institute
TU Shannon: Midlands Midwest
Athlone, Ireland
blee@ait.ie

Visual surveillance systems have been improving rapidly over the recent past, becoming more capable and pervasive with incorporation of artificial intelligence. At the same time such surveillance systems are exposing the public to new privacy and security threats. There have been an increasing number of reports of blatant abuse of surveillance technologies. To counteract this, data privacy regulations (e.g. GDPR in Europe) have provided guidelines for data collection and data processing. However, there is still a need for a private and secure method of model training for advanced machine learning and deep learning algorithms. To this end, in this paper we propose a privacy-preserved method for visual surveillance. We first develop a dataset of privacy preserved videos. The data in these videos is masked using Gaussian Mixture Model (GMM) and selective encryption. We then train high-performance object detection models on the generated dataset. The proposed method utilizes state-of-art object detection deep learning models (viz. YOLO*v4* and YOLO*v5*) to perform human/object detection in masked videos. The results are encouraging, and are pointers to the viability of the use of modern day deep learning models for object detection in privacy-preserved videos.

CCTV Videos, GDPR, Object Detection, Privacy Preserving Surveillance, YOLO

1. INTRODUCTION

Global technology landscape is evolving expeditiously. Legacy approaches and systems are being by replaced by modern services and solutions creating new business models and use cases. Innovations in small scale chip design have produced incremental size reduction and enhanced performance of sensors and other electronics gadgets, for example surveillance cameras. These surveillance cameras are employed for monitoring in different sectors. For instance, in healthcare settings, due to shortage of skilled care staff, cameras are used for monitoring and providing timely care to elderly (Cournan et al. 2018). Another important use-case of video cameras is for security and safety purposes. CCTV (closed-circuit television) cameras are used for round the clock look-out and security at safety critical infrastructure such as banks, ATMs, airports, sensitive and non-sensitive government installations. Overhead cameras alongside roads are used for traffic control, plate recognition and to provide a better road experience (Semertzidis et al. 2010). Security cameras are becoming ubiquitous in our lives. According an estimate, there are one billion surveillance cameras in operation till 2021 (Cosgrove 2019).

The pervasive surveillance technologies are exposing general public to various privacy and security vulnerabilities. Report published by a privacy watchdog has reported that an average Londoner is caught by 300 cameras in one day (Pillai 2012). Four council staff members were reported for voyeurism through CCTV cameras (BBC News 06 Dec,2005) in UK. In similar incident in USA, a technician hacked a surveillance camera system to exploit women when they were in vulnerable circumstances (Geiger 2021). State officials are also not immune from privacy breaches. Former German Chancellor Angela Merkel was also fell victim to privacy intrusion when a museum guard directed a PTZ

(pan-tilt-zoom) camera towards the chancellor's residence's window (Furlong 2006). In an attack on the surveillance company Verkada, around 150,000 cameras were hacked, which left Tesla factories, psychiatric hospitals and schools exposed to outsiders (Meye 2021). Surveillance cameras are also being weaponized by authoritarian regimes to control free speech (Richards 2013). The recurrent theme in these incidents is the exploitation of surveillance technology to attack victim's privacy and security. To counteract such threats, the European Parliament passed General Data Protection Regulation (GDPR), which enforces privacy protection of individuals, objects and locations in recorded videos (Asghar et al. 2019). This situation warrants development of video-based human activity recognition solutions with privacy-preserving capabilities and revealing minimum identifiable information about the person under surveillance.

In recent times, deep learning based computer vision methods have been witnessing tremendous success at identifying objects and actions in videos. Deep learning based surveillance solutions are already being used. In fact, its due to the ability of these smart solutions of learning minute details, and then quickly building meaningful information from them, which is quite worrisome. These concerns necessitates development of surveillance system which can perform its task using privacy-preserving videos.

Under GDPR law (Asghar et al. 2019) it is necessary that operators who monitor surveillance systems or who interact with video data should have limited access to visually identifiable features of persons. The main of contribution of this paper is method to develop privacy preserved video data for training and testing the deep learning models. Also application of same privacy preserving methods in surveillance systems so that surveillance system operator have limited or law compliant access to video data.

In this research article, we present a proof-of-concept of a privacy-preserved surveillance system. The proposed method is based on modern state-of-art object detection deep learning models YOLOv4 and YOLOv5. These models are trained on masked videos. The goal of masking videos is to ensure privacy requirements are fulfilled.

This research paper is structured into five sections. Section II presents a concise overview of related works. Section III presents details of the proposed methodology where we highlight our suggested solution. Section IV comprises of the experiments and results, where we discuss the implementation and results of our proposed solution. Lastly, Section-V contains some concluding remarks.

2. RELATED WORK

GDPR regulations define various categories of spaces for visual data privacy compliance. This helps identifying grey ares where a privacy preserving solution is required. This grey area includes spaces which are at the junction of public and privates spaces. Figure1 illustrates this concept comprehensively. There are private spaces which one would want to protect and monitor, but at the same these personal spaces share a common boundary with a region of public (impersonal) spaces. Such public spaces called 'protected spaces'. An individual is very much within their rights to record his personal space, similarly an individual in protected space also has right to the preservation of his data.

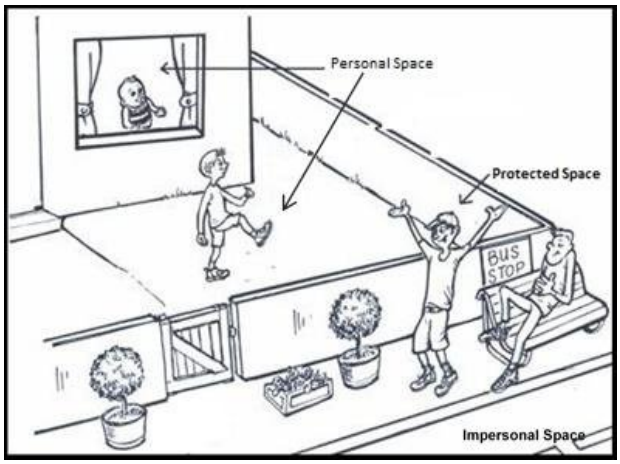

Figure 1: *Identification of personal, protected, and impersonal space in the context of privacy-preservation regulations (Buyya et al. 2009; Asghar et al. 2019)*

Figure 2: *Example of the output of a Live Facial Recognition system (Humphrys 16 August, 2019)*

Moreover, an individual in his personal space can also share data with third parties such as police. Hence, these spaces require privacy-preserving surveillance solutions. In its article 25, GDPR also introduces two design paradigms, namely Data protection-by-Design and Data protection-by-Default. Data protection-by-Default restricts the

collection of data. It enforces that only necessary data or information about a person is collected, and nothing else. Data Protection-by-Design means the software developer or service provider will provide built-in data privacy safeguards through technological means. The regulation has made it mandatory but left to choice of solution developer to decide which method to use. A common visual data privacy protection method is redaction of data. The redaction of visual data involves hiding or masking of visually identifiable features of a said individual in video or it can be applied to entire video. This would protect privacy of all objects, background and location in video. After redaction, computer vision (CV) methods are used for object detection and object tracking to detect or track objects in the video. GDPR has greatly emphasized visual privacy methods, pseudonymisation and anonymization. Pseudonymisation hides information partially and utilizes reversible methods such as tokenization, scrambling, use of wrong colors etc., whereas anonymization completely hides the information of the subject and is typically irreversible (e.g. blurring, mosaic, pixelation, masking, wrapping etc.)

2.1. Privacy Preserving in Videos

A typical image as acquired from a surveillance system, and passed through a deep learning based detection system, is shown in Fig. 2. It can be seen this image is clearly providing excessive amounts of information about the persons in frame, thereby violating their privacy. To combat such threats, Ivasic-Kos et al. have used 2D Gaussian filtering to blur out the human body silhouettes that otherwise expose identity information about the person (Ivasic-Kos et al. 2014). Liu et al. proposed a visually privacy-preserving fall detection framework for elderly care (Liu et al. 2021). They achieved privacy through visual shielding using compressive sensing, however the complexity of system makes it unfit for real-time use. Ryoo et al. proposed activity recognition system based on transforming high resolution videos to extreme low resolution videos (Ryoo et al. 2017). A classifier later classifies the actions performed in video. The pitfall of this method is downgrading videos from high resolution to very low resolution. Tariq et al. proposed a multi level access control scheme for surveillance systems (Tariq et al. 2020). The scheme has three tiers with restricted access to video frames. Although this scheme limits access to video data but it is not robust enough for real-time applications. Fitwi et al. proposed a selective frame encryption scheme to protect the privacy of general public (Fitwi and Chen 2020). The proposed scheme encrypts the video at the edge node itself using lightweight dynamic chaotic image enciphering before transferring them to the fog/cloud server. The frames are deciphered at the server and only the frames with violent or suspicious behaviour are stored. Authors in (Fitwi et al. 2020) proposed a framework which scrambles the video frames at edge using Reversible Chaotic Masking (ReCAM) coupled with a simple foreground object detector to discount frames with no information. Miran et al. presented a fully homomorphic encryption (FHE) based scheme for human action recognition for old-age care homes Kim et al. (2021). They achieved this by first acquiring action information from the skeleton joints and then encrypting the information acquired. A semantic segmentation based masking framework with different access levels is presented in (Abbasi et al. 2019). While the access control can provide some protection, the semantic segmentation applied in not good enough at hiding visual features of subjects.

To summarize, there is a need for a high-performance and robust privacy-preserving human detection system which can be used in modern day surveillance set-ups to facilitate optimal levels of privacy and obfuscation for the subjects recorded in the videos. This paper attempts to present one viable solution.

3. PROPOSED METHODOLOGY

3.1. System Design

In order to develop a smart surveillance system for privacy preserved videos, our proposed system model is presented in Figure 3. The different steps adopted in this design are listed below:

1. The camera captures the video of a person.
2. The camera will have an on-board processing unit which will perform the masking of videos, or alternatively the videos are transferred to the fog/cloud node where masking would to be performed.
3. After masking the videos are transferred to the deep learning based object detection model.
4. In the last step, the model performs classification of masked videos and sends the results to the control room.

3.2. Privacy Preserved videos

For leveraging any deep learning model, the primary requirement is an optimal dataset. There exist many video datasets, however there are not privacy preserved. Therefore, we decided to develop our own privacy-preserved surveillance video dataset. There were some good candidates for this but they were removed from public domains by their

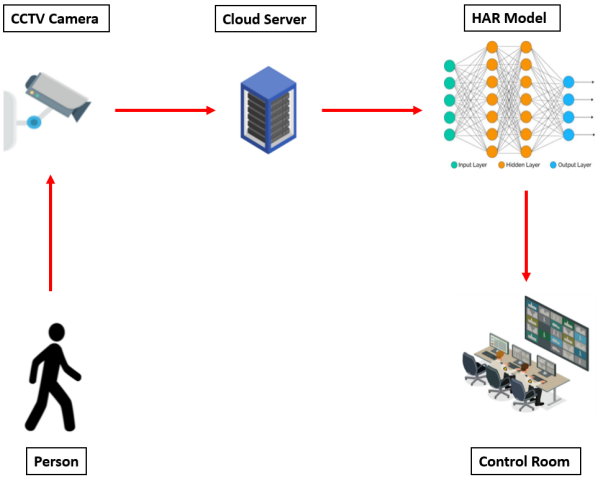

Figure 3: *Proposed Privacy-Preserving Human Detection Framework using a Deep Learning Model*

publishers as they were over privacy and security concerns (Harvey 2021). To solve this problem, we collected publicly available surveillance videos from different sources and also recorded our own. The videos collected were of different resolutions. For achieving a uniformity of dataset, first all videos were masked. The privacy preserving of videos is achieved through extracting the regions of interest (ROI) and then performing selective encryption (SE) using reverse XOR. Gaussian Mixture Model (GMM) is employed to extract the ROI mask of each frame using background modelling. The SE in combination reverse XOR is used to make the video GDPR compliant i.e. hiding the identity of person in video. After masking their frames were extracted and were resized to 640×480 pixels resolution. Table 1 provides details of dataset. Images in Figure 4 show some samples of unmasked video frames (a-d) and their corresponding masked frames (e-h).

Table 1: *Dataset details*

Dataset	No. of Images	Resolution
Training	400	640×480
Testing	100	640×480

3.3. Object Detection Models

The research area of object detection in digital images using deep learning approaches has seen enormous strides during the recent past (Zaidi et al. 2022). With new and improved models being released frequently, the use of such improved models in different avenues is increasing significantly (Aslam and Curry 2021). The technical literature in now replete with a variety of object detection models which are available for researchers and practitioners to choose from – based on the specific requirements of speed, parameter count,

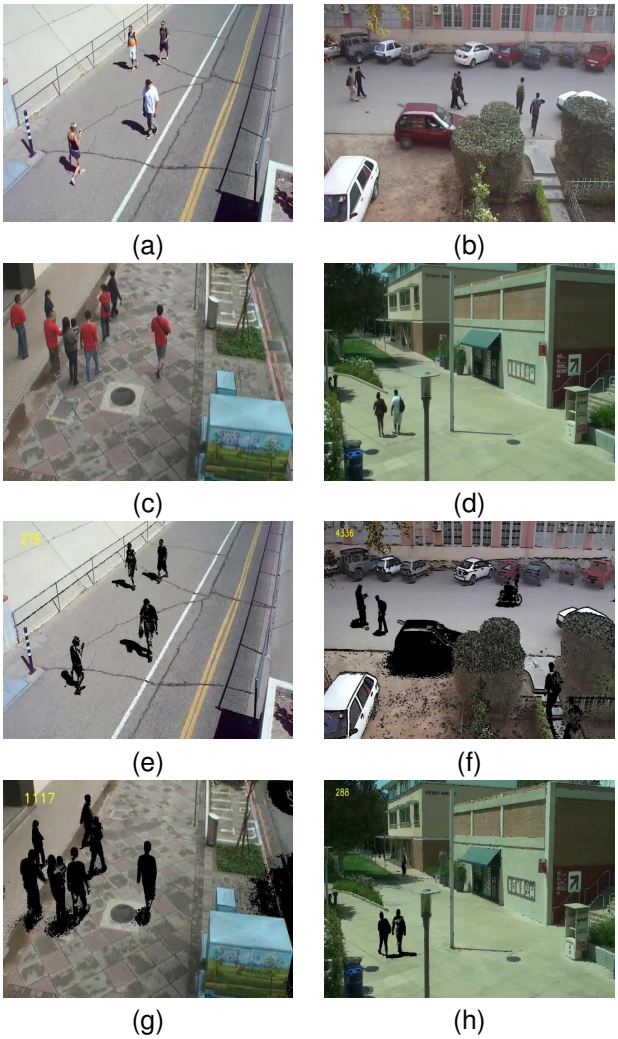

Figure 4: *Some Samples of Unmasked (a-d) and Maksed (e-h) Video Dataset Frames*

and performance. Amongst the available models, the class of models referred to as You-Only-Look-Once (YOLO) have been enormously successful due to a wide range of advantages that they offer (Zaidi et al. 2022). This work therefore utilizes the recent YOLOv4 and YOLOv5 object detection models.

3.3.1. YOLOv4

YOLOv4 was presented by Alexey Bochkovskiy (Bochkovskiy et al. 2020), and is based on Darknet model akin to its predecessors (Redmon et al. 2015). The training and inference performance of YOLOv4 was improved by adopting Bag of freebies (BoF) and Bag of specials (BoS) strategies. Bag of freebies (BoF) is a collection of methods that increase training cost without affecting inference cost. In YOLOv4 BoF methods data augmentation, class label smoothing and objective function of BBox regression are used. In contrast to BoF methods, Bag of specials (BoS) methods only affect inference cost. This effect may be small but can significantly improve object detection. Bag of specials (BoS) methods deployed

Table 2: *YOLOv4 vs YOLOv5 comparison*

Model	YOLOv4	YOLOv5
Neural Network Type	Fully convolution	Fully convolution
Backbone Feature Extractor	CSPDarknet53	CSPDarknet53
Neck	SPP and PANet	PANet
Head	YOLO layer	YOLO layer

Table 3: *Object Detection models training summary*

Model	epochs	batch size	training time (hours)
YOLOv4	1000	2	3.5
YOLOv5s	30	2	0.567
YOLOv45m	30	2	1.281
YOLOv5l	30	2	1.904

Table 4: *Object detection models comparison*

Model	mAP	Precision	Recall	F1 score	Inference time (ms)	Parameters
YOLOv4	0.60	0.84	0.96	0.896	144.07	$\sim$60 million
YOLOv5s	0.967	0.974	0.875	0.921	82.5	7.2 million
YOLOv45m	0.945	0.94	0.898	0.918	103.93	21.2 million
YOLOv5l	0.944	0.929	0.886	0.906	225.5	46.5 million

in YOLOv4 are Mish activation, Cross-stage partial connections (CSP) and Multi-input weighted residual connections (MiWRC). These additions improved the accuracy of model without necessarily affecting the inference time but only increment in the training cost.

3.3.2. YOLOv5

YOLOv5 was made publicly available in 2020 (Jocher 2020). It shares its architectural details with YOLOv4. Its backbone is based on CSPDarknet53 which helps with the repetitive gradient information in large backbones and combines gradient variations into feature map that decreases the inference speed, improves accuracy, and shrinks the model size by lowering number of the parameters. In its neck part, YOLOv5 uses PANet (path aggregation network), which acts as feature pyramid network (FPN) and increases detection of low-level features. Table 2 shows a summary of the architectural design comparison of YOLOv4 ad YOLOv5.

4. EXPERIMENT AND RESULTS

We trained and tested our model on Nvidia A100-SXM4-40GB GPUs and also Google Colab with Tesla K80 GPUs. Table 3 provides summary of training parameters and training times for different models. Figure 5 shows the obtained detection results from our models. It can be seen that our models have successfully detected the subjects in masked frames with high accuracy.

Table 4 compares different performance metrics for object detection models. It needs to be mentioned that YOLOv4 needed to be trained to a much higher number of epochs (1000) as compared to the YOLOv5 variants (s,m,l – 30 epochs each) to achieve the same level of accuracy and loss parameters during training. This clearly has an effect on the overall training time of the different models, and YOLOv5 excels by providing a very short training time for the exact same training dataset. Table 4 on the other hand shows the inference time taken by the different models for performing object detection on a sample image from the testing dataset. YOLOv5s excels here as expected due to the very low parameter count. An interesting observation arising out of Table 4 is that YOLOv5l required a higher inference time than the previous generation YOLOv4, although YOLOv5l does provide higher values of mAP, Precision, Recall, and F1 Score (as compared to YOLOv4) which is expected from a newer generation model.

A deep learning model is considered robust and optimal when there is a balance between Precision and Recall (assuming both are high), and F1 score is a good indication of this balance. It can be observed from the data in Table 4 that Precision and Recall for both models are well balanced for our custom dataset. However in terms of mean average precision (mAP), YOLOv5 models perform better than YOLOv4. Table 4 also compares inference times of different models. It can be concluded from Table 4 that YOLOv5s model is most optimum model to used for this human detection as it has highest mAP and minimum training and inference times. In fact, for this particular use-case, the YOLOv5s

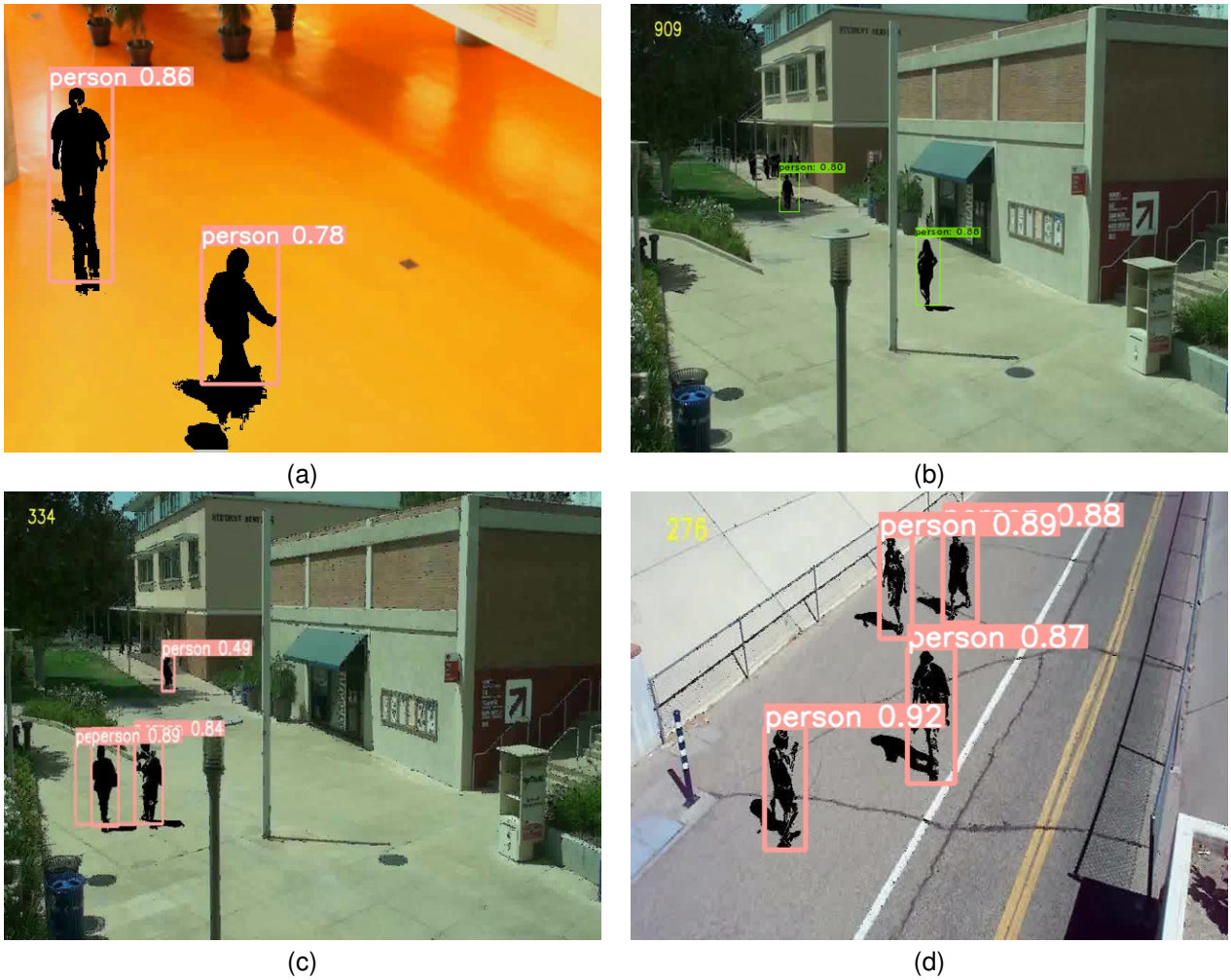

Figure 5: *Detection results obtained using the deep learning models over the custom-built privacy-preserving dataset*

model outperforms the full-scale elaborate YOLOv5l model which is a surprising result. With such low inference times and the low parameter count, YOLOv5s becomes the ideal candidate for use in resource-constrained nodes.

5. CONCLUSION AND FUTURE WORK

The smart visual surveillance systems are pervasive in our daily lives. They provide us with sense of security, control and comfort. But at the same time they are leaving us vulnerable to numerous privacy threats. Governments have passed laws to restrict access and processing of data collected through the solution. However, there remains a gap for a private and secure method of training for advanced machine learning and deep learning algorithms.

In this article, we have proposed a privacy preserving smart visual surveillance solution. Our proposed solution uses state of the art object models (YOLOv4, YOLOv5) in tandem with video masking methods to perform the required task. So far we have only performed object detection in our proposed system. As an avenue for future research, this work can be extended to perform person tracking in real-time videos. The low parameter count and very low inference time of YOLOv5s is expected to be of significant advantage in that regard.

REFERENCES

Abbasi, M. H., Majidi, B., Eshghi, M. and Abbasi, E. H. (2019), Deep visual privacy preserving for internet of robotic things, *in* '2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)', pp. 292–296.

Asghar, M. N., Kanwal, N., Lee, B., Fleury, M., Herbst, M. and Qiao, Y. (2019), 'Visual surveillance within the eu general data protection regulation: A technology perspective', *IEEE Access* **7**, 111709–111726.

Aslam, A. and Curry, E. (2021), 'A survey on object detection for the internet of multimedia

things (iomt) using deep learning and event-based middleware: approaches, challenges, and future directions', *Image and Vision Computing* **106**, 104095.

BBC News (06 Dec,2005), 'CCTV staff 'spied on naked woman", *BBC News* .
URL: *http://news.bbc.co.uk/2/hi/uk_news/england/merseyside/4503244.stm*

Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y. M. (2020), 'Yolov4: Optimal speed and accuracy of object detection', *ArXiv* **abs/2004.10934**.

Buyya, R., Selvi, S. T. and Chu, X. (2009), *Object-oriented programming with Java: essentials and applications*, Tata McGraw-Hill.

Cosgrove, E. (2019), 'One billion surveillance cameras will be watching around the world in 2021, a new study says', *CNBC* .
URL: *https://www.cnbc.com/2019/12/06/one-billion-surveillance-cameras-will-be-watching-globally-in-2021.html*

Cournan, M., Fusco-Gessick, B. and Wright, L. (2018), 'Improving patient safety through video monitoring', *Rehabilitation Nursing* **43**, 111–115.

Fitwi, A. and Chen, Y. (2020), Privacy-preserving selective video surveillance, *in* '2020 29th International Conference on Computer Communications and Networks (ICCCN)', pp. 1–10.

Fitwi, A., Chen, Y. and Zhu, S. (2020), Prise: Slenderized privacy-preserving surveillance as an edge service, *in* '2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)', pp. 125–134.

Furlong, R. (2006), 'Germans probe merkel spy camera', *BBC News* .
URL: *http://news.bbc.co.uk/2/hi/europe/4849806.stm*

Geiger, D. (2021), 'ADT technician hacked security cameras at hundreds of homes, spied on 'attractive' women and couples', *Oxygen-True Crime* .
URL: *https://www.oxygen.com/crime-news/telesforo-aviles-admits-to-hacking-adt-home-security-cameras*

Harvey, Adam. LaPlace, J. (2021), 'Exposing.ai'.
URL: *https://exposing.ai/duke_mtmc/*

Humphrys, J. (16 August, 2019), 'John humphrys - facial recognition cameras: How worried should we be?', *youGov* .
URL: *https://yougov.co.uk/topics/politics/articles-reports/2019/08/16/john-humphrys-facial-recognition-cameras-how-worri*

Ivasic-Kos, M., Iosifidis, A., Tefas, A. and Pitas, I. (2014), Person de-identification in activity videos, *in* '2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)', pp. 1294–1299.

Jocher, G. (2020), 'Yolov5', *GitHub* .
URL: *https://github.com/ultralytics/yolov5*

Kim, M., Jiang, X., Lauter, K., Ismayilzada, E. and Shams, S. (2021), 'Hear: Human action recognition via neural networks on homomorphically encrypted data'.

Liu, J., Tan, R., Han, G., Sun, N. and Kwong, S. (2021), 'Privacy-preserving in-home fall detection using visual shielding sensing and private information-embedding', *IEEE Transactions on Multimedia* **23**, 3684–3699.

Meye, C. (2021), 'Breach of 150,000 surveillance cameras sparks credential concerns', *BBC News* .
URL: *https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/march/breach-surveillance-cameras-sparks-credential-concerns*

Pillai, G. (2012), 'Caught on camera: You are filmed on CCTV 300 times a day in london', *International Business Times* .
URL: *https://www.ibtimes.co.uk/britain-cctv-camera-surveillance-watch-london-big-312382*

Redmon, J., Divvala, S. K., Girshick, R. B. and Farhadi, A. (2015), 'You only look once: Unified, real-time object detection', *CoRR* **abs/1506.02640**.

Richards, N. M. (2013), 'The Dangers of Surveillance', *Harvard Law Review* **127**, 1934–1965.

Ryoo, M. S., Rothrock, B., Fleming, C. and Yang, H. J. (2017), Privacy-preserving human activity recognition from extreme low resolution, *in* 'Thirty-First AAAI Conference on Artificial Intelligence'.

Semertzidis, T., K.Dimitropoulos, A.Koutsia and N.Grammalidis (2010), 'Video sensor network for real-time traffic monitoring and surveillance', *IET Intelligent Transport Systems* **4**, 103–112(9).

Tariq, F., Kanwal, N., Ansari, M. S., Afzaal, A., Asghar, M. N. and Anjum, M. J. (2020), Towards a privacy preserving surveillance approach for smart cities, *in* '3rd Smart Cities Symposium (SCS 2020)', Vol. 2020, pp. 450–455.

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M. and Lee, B. (2022), 'A survey of modern deep learning based object detection models', *Digital Signal Processing* **126**, 103514.