Computing and Informatics, Vol. 44, 2025, 1273–1294, doi: 10.31577/cai_2025_6_1273

DEPTH-WISE AND DEPTH-WISE SEPARABLE YOLO MODELS FOR CONCEALED OBJECT DETECTION USING TERAHERTZ IMAGES

Singara Singh Kasana, Lakshmy Santhosh

Computer Science and Information Technology
Central University of Haryana
Mahendergarh, 123031
Haryana, India
e-mail: singarasingh@cuh.ac.in, lakshmys00@gmail.com

Abstract. Terahertz imaging is highly effective for detecting concealed objects due to its non-harming nature and its ability to penetrate materials like clothes, paper and plastic, etc. In environment where detection technologies are limited, terahertz imaging emerges as one of the effective and safest methods available. Unlike techniques such as X-rays, it does not emit harmful radiation, making it suitable for surveillance applications. However, many existing object detection models are computationally intensive which can hinder their deployment in real time or resource constrained environments. To address this issue, traditional convolutional operations in the deep learning models have been replaced with depth-wise convolutions and depth-wise separable convolutions in proposed approach. These modifications significantly reduce the number of trainable parameters and computational load during model training. The optimized architecture has been integrated into widely used object detection models – namely YOLOv5m and YOLOv8m, using terahertz images of concealed objects as input. This integration enhances training efficiency with minimal loss in accuracy, making the models more suitable for deployment on devices with limited computational power and memory.

Keywords: Depth-wise separable convolution, terahertz, YOLO, convolution neural networks, mAP, precision, recall

Mathematics Subject Classification 2010: 68T45

1 INTRODUCTION

Object detection is a critical component of computer vision, which focuses on identifying and localizing objects within images or video. The primary objective is to classify objects and provide their precise locations, typically using bounding boxes. Concealed Object Detection, on the other hand, is a specialized subfield of object detection which focuses specifically on identifying hidden objects. This domain is particularly relevant in security and surveillance contexts, where detecting concealed threats – such as weapons, explosives, or other illicit items – is critical. Techniques employed in concealed object detection often require advanced imaging methods, such as terahertz (THz) imaging, millimeter-wave imaging, or X-ray systems, which can penetrate materials and reveal objects that would otherwise be hidden from conventional visual detection. It plays a pivotal role in enhancing security measures in various real-time application, including airport screening, public surveillance, military applications, etc.

The use of THz images for concealed object detection is preferred for the following reasons:

- Terahertz radiation easily passes through most non-metallic and non-polar materials, allowing THz systems to "see through" barriers like packaging, corrugated cardboard, clothing, shoes, and book bags. This capability enables the detection of potentially dangerous materials hidden within these items.
- Many materials relevant to security applications, such as explosives, chemical agents, and biological agents, exhibit distinct THz spectra that can be utilized to fingerprint and identify these concealed substances.
- Terahertz radiation presents either no health risk or minimal risk to both the individual being scanned by a THz system and the system operator.

Terahertz images with some hidden objects are shown in Figure 1. In these images, the blue and black areas represent the hidden objects while the red and yellow areas indicate the body of the person. The variation in frequency between the body and the concealed objects helps in their detecting through terahertz images.

There are two types of THz imaging:

Active THz Imaging: It refers to a technique that utilizes THz radiation to capture detailed images of objects or materials. It involves the use of a THz source to illuminate the target. This approach allows for enhanced contrast and sensitivity, making it particularly effective for detecting concealed objects and identifying materials.

Passive THz Imaging: It captures images using THz radiation emitted naturally by objects, rather than using an external source of THz waves. This method relies on the thermal radiation emitted by materials, which can be detected and converted into images.

Traditionally, object detection relied on feature-based methods, which required extensive manual feature extraction and often struggled with varying lighting conditions, occlusions, and complex backgrounds. However, the advent of deep learning has revolutionized this field, enabling more accurate and efficient detection through the use of Convolutional Neural Networks (CNNs), You Only Look Once (YOLO), Faster R-CNN, and Single Shot MultiBox Detector (SSD) are leading the way in both accuracy and speed. These models automatically learn hierarchical features from the input image/video, significantly improving performance over conventional techniques. The YOLO framework processes images in a single pass, allowing for rapid localization and classification of multiple objects within an image. The speed provided by the YOLO model is particularly beneficial in security applications where timely responses are essential. YOLO's architecture divides the image into a grid and assigns bounding boxes to each grid cell, predicting the likelihood of various objects appearing in those boxes. This unique approach enables the model to effectively handle overlapping objects and complex scenes, making it suitable for detecting concealed items in various environments. Specifically, *YOLOv5* and its successor, *YOLOv8*, have demonstrated strong capabilities in detecting objects with minimal latency, making them suitable candidates for concealed object detection tasks.

This research explores the integration of depth-wise and depth-wise separable convolutions into the YOLO framework. Depth-wise convolution techniques have gained significant attention in recent years for their ability to improve model efficiency without compromising accuracy [1]. By replacing traditional convolutional layers with depth-wise convolution, we aim to enhance the feature extraction process within the YOLO model, potentially improving its performance in object detection tasks. Depth-wise separable convolution decomposes the convolution operation into depth-wise and point-wise convolutions, which reduces computational complexity [2]. These convolutional techniques not only reduce the number of trainable parameters but also enhance computational efficiency, making the models more accessible for deployment on devices with limited processing power. By leveraging these advancements, the study aims to improve the detection accuracy and speed of YOLOv5 and YOLOv8 models when applied to terahertz imaging for concealed object detection. The findings from this research are expected to contribute to the development of more robust surveillance systems capable of operating effectively in diverse environments.

The paper is structured as follows: Section 2 covers the analysis of existing concealed object detection techniques; Section 3 outlines the proposed technique; Section 4 presents the experimental results and analysis. The paper concludes with Section 5.

2 RELATED WORKS

In this section, we review the existing concealed object detection techniques based on terahertz images. These techniques are classified into three categories as discussed.

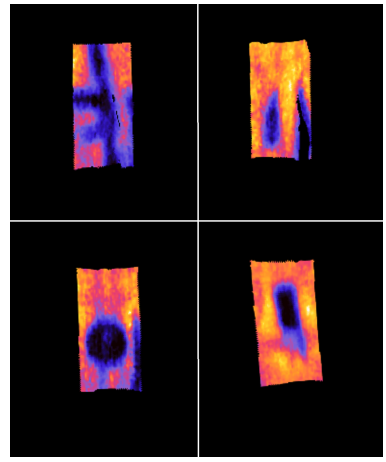


Figure 1. Some of terahertz images with hidden objects

2.1 Transformer Based Concealed Object Detection Techniques

Transformer-based techniques for concealed object detection have gained prominence due to their ability to effectively analyze complex visual data. These techniques leverage the transformer architecture, originally designed for natural language processing, which uses self-attention mechanisms to capture long-range dependencies and contextual relationships within images. In concealed object detection, transformers can effectively highlight relevant features and distinguish hidden objects, such as weapons or contraband, from the background. This is achieved through models that process spatial information in parallel, allowing for greater efficiency and accuracy compared to traditional CNNs.

Recent advancements have shown that integrating transformers with multimodal data, such as THz imaging, significantly enhances detection capabilities. For example, transformers can combine features from different imaging modalities to provide a more comprehensive understanding of the scene, thereby improving detection rates. For instance, [3] demonstrated the advantages of transformer models in detecting concealed items in security applications, showcasing improvements in accuracy and processing speed compared to conventional methods. Additionally, [4] explored the integration of transformers with THz imaging to enhance object recognition in challenging conditions.

2.2 Attention Based Concealed Object Detection Techniques

Attention-based concealed object detection techniques leverage attention mechanisms to improve the identification of hidden objects in various imaging modalities, including traditional and advanced imaging systems. By focusing on the most rele-

vant parts of an image, these techniques enhance model performance in challenging detection tasks, such as identifying concealed weapons or contraband in security contexts.

Attention mechanisms, commonly used in deep learning architectures, allow models to weigh the significance of different regions within an image. This capability is particularly advantageous in complex environments where background noise can obscure concealed objects. By dynamically adjusting focus, attention-based models can isolate features that indicate the presence of hidden items, leading to higher detection accuracy and reduced false positives.

Cheng and Lucyszyn [5] presented a novel approach for detecting concealed objects with limited labelled data. By leveraging few-shot learning and an improved pseudo-annotation mechanism, the method enhances model generalization and accuracy in sub-terahertz imaging. This work significantly advances data-efficient detection for security screening applications.

Cheng et al. [6] proposed Adaptation-YOLO, an advanced detection framework for active THz security images, integrating an Adaptive Context-Aware Attention Network (ACAN) to improve the detection of concealed objects. The approach enhances feature representation and accurately identifies low-contrast, subtle objects, achieving superior performance compared to baseline YOLO models. This work demonstrates a practical and efficient solution for real-time security and surveillance applications.

The adaptability of attention-based methods makes them suitable for a variety of applications beyond security, including medical imaging and industrial inspection. As these techniques continue to evolve, they promise to offer even greater accuracy and efficiency in detecting concealed objects across multiple domains.

2.3 YOLO Based Concealed Object Detection Techniques

In the field of concealed object detection, terahertz security screening cameras achieve a balance between low radiation exposure and efficient detection, generally providing only approximate object locations. Yang et al. [7] proposed a CNN based technique that employs sparse and low-rank decomposition for detecting objects in THz images. Detailed recognition is achieved through supervised training using the Faster R-CNN model.

Wang et al. [8] proposed a normalized accumulation map-based training approach for concealed object detection in millimeter-wave images. This approach, when integrated with the YOLOv2 model, an improvement in mean Average Precision (mAP) was observed. In [9], Wang et al. proposed a self-paced feature fusion network for concealed Active Millimeter Wave (AMMW) and Passive Millimeter Wave (PMMW) images.

Kowalski [10] compared concealed object detection techniques using various types of clothing with passive imagers in the terahertz range and mid-wavelength infrared ranges. The models used in this study include YOLOv3 and Region-based Fully Convolutional Networks (R-FCN). A real-time detection method for identify-

ing concealed metallic weapons on the human body using PMMW imaging based on YOLOv3 was proposed by Pang et al. [11].

Danso et al. [12] developed a hidden object detection model using terahertz images. Their model is an improved version of YOLOv5 with BiFPN at the neck of YOLOv5 to improve performance at low resolutions. They also used transfer learning by fine-tuning the pre-training weights of the backbone for effective migration learning. Similarly, Jayachitra et al. [13] proposed a YOLOv5 model integrated with a novel mutation-enabled salp swarm algorithm for parameter optimization and model fine-tuning.

Xu et al. [14] proposed a specialized deep learning network for the automatic and accurate real-time detection of objects in passive THz images. Bilateral filter integrated in CNN, Multi-scale Filtering and Geometric augmentation and an improved YOLOv5l is used in this model. Ge et al. [15] proposed preprocessing techniques for THz images using Non-Local Mean (NLM) filtering and histogram equalization prior to object detection with YOLOv7. Their method using NLM filtering achieved the highest detection accuracy.

Depth-wise and Depth-wise separable convolutions are two modified versions of convolution operation that can reduce the number of trainable parameters and the complexity of a model without significantly sacrificing accuracy. Depth-wise separable convolution can be integrity into various models in place of standard convolution [16].

Panigrahi and Raju [17] applied depth-wise separable convolution to enhance YOLOv2, resulting in the DSM-IDM-YOLO model. Their framework is computationally efficient due to its convolution design and a moderate number of layers, aiming to improve performance while minimizing computational overhead. This work was conducted on a pedestrian dataset.

Qin et al. [18] proposed a method combining a classification model with a detection model for fire detection. Initially, depth-wise separable convolution is used to classify fire images. For images classified as containing fire, the YOLOv3 target regression function is then utilized to output the fire position information. Training was performed on a public dataset, achieving a detection accuracy of 98 % and a detection rate of 38 frames per second.

Liu et al. [19] developed a sea surface object detection algorithm based on YOLOv4, integrating reverse depth-wise separable convolution into both the backbone network and feature fusion network of YOLOv4. This approach reduced the number of weights by 40 %, enhanced detection speed by over 20 %, and slightly improved mAP in the datasets used.

2.4 Limitations of Existing Models

Despite extensive research in the field of concealed object detection using YOLO models, transformer models and other attentions based CNN models, there are certain limitations.

Lack of Training Data: Deep learning models require extensive training data for optimal performance. One major problem is the unavailability or limited availability of terahertz and active millimeter wave datasets.

Complexity and Computational Demands: Many deep learning approaches have high computational complexity, which limits their ability to perform real-time detection in resource-constrained environments.

Data Quality and Enhancement: Commonly used datasets such as terahertz and infrared images, often suffer from low quality. Enhancing the quality of these images by using existing techniques is challenging.

Challenges with Small Object Detection: Deep learning based techniques struggle to accurately detect small concealed objects which impacts overall detection performance.

Difficulty in Determining Precise Object Location: Achieving precise object localization is a critical challenge for detection models, especially due to noise and clutter in images like terahertz.

2.5 Motivations of the Proposed Work

The following are the motivational points to carry out the proposed research:

- Concealed object detection is crucial for security surveillance, as safety in public places is a significant concern. Terahertz (THz) waves, which fall between the infrared and microwave frequency ranges, possess the unique ability to penetrate opaque materials such as clothing or packaging. This characteristic enables them to detect objects with different frequencies. Despite the potential of terahertz imaging to significantly improve public security, research in this area remains limited.
- YOLO models are capable of providing state-of-the-art performance in object detection and image segmentation tasks. They offer superior speed and accuracy compared to traditional object detection models. Given that both speed and accuracy are crucial in object detection in surveillance systems, YOLO models are an excellent choice.
- Due to the potential of terahertz imaging and the performance of YOLO models, integrating these technologies and optimizing results through architectural modifications can be highly beneficial. Utilizing terahertz imaging and YOLO models can significantly enhance security systems' ability to detect threats effectively, ultimately creating safer environments for everyone.

2.6 Contribution of the Proposed Work

The contributions of the proposed work are as follows:

- Since YOLO models perform well on images of different frequencies like terahertz, millimeter waves, etc., we use YOLO models for creating a faster concealed object detection model.
- For the proposed model, we used two versions of YOLO (YOLOv5 and YOLOv8). Using the medium versions of both these algorithms and replacing convolutional layers by depth-wise and depth-wise separable convolutional layers, we are able to get good results in terms of mAP, precision and recall.

3 PROPOSED TECHNIQUES

3.1 DW-YOLOv5 and DWS-YOLOv5

YOLOv5 is a successor of previous versions of YOLO models, enhancing both architecture and performance. It is a single-stage object detection pipeline capable of real-time inference on various devices. It has different version like s, m and l, each varying in speed and number of parameters. In this research, we use the YOLOv5m architecture. Its architecture contains three main sections: the backbone, neck and head, as shown in Figure 2.

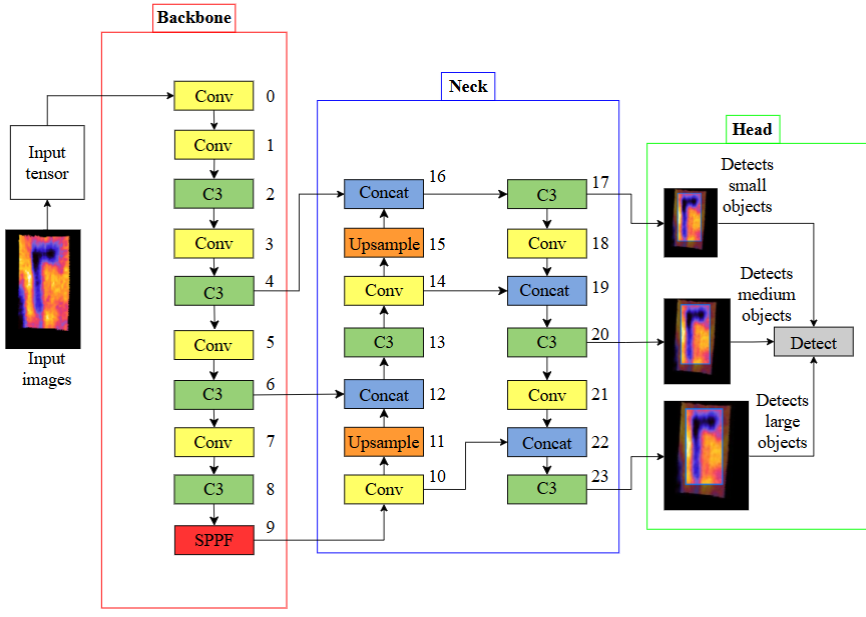


Figure 2. YOLOv5m architecture

Figure 2 illustrates the architecture of YOLOv5m, featuring various blocks such as Conv, C3, Concat, Upsample and Detect. The Detect block, which receives input from the C3 block (17) detects smaller objects, whereas the C3 block (20) detects medium-sized objects and the C3 block (23) detects larger objects. The structure of C3 contains 3 Conv blocks, bottleneck layers and Concat whereas the structure of SPPF contains Conv blocks, maxpooling layers and Concat.

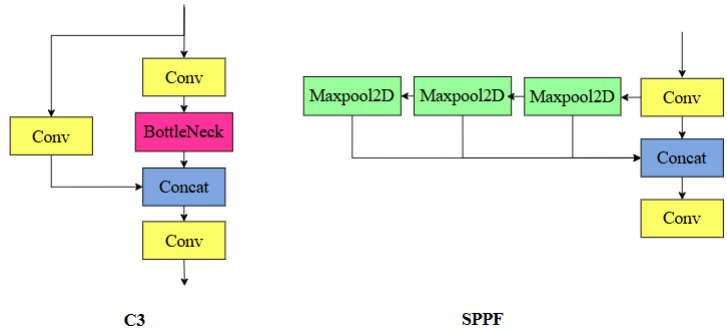


Figure 3. C3 and SPPF structure

Backbone: The backbone is responsible for feature extraction. Input image is converted into tensors which are then sent to the backbone for feature extraction. The backbone is composed of CSPDarknet53 and includes multiple CBS and C3 modules, culminated in an SPPF module. CBS (Conv2d+BatchNormalization+ SiLU) layer is shown as 'Conv' in Figure 2. The CBS module supports C3 module for feature extraction and SPPF module enhances the feature expression of the backbone, as shown in Figure 3.

Neck: The neck consists of Path Aggregation Network (PANet). PANet concatenates features from the backbone with its own output. These combined features undergo convolution layers to reduce computational complexity. PANet includes convolutional layers, BatchNormalization, and SiLU activation, along with upsampling and downsampling operations. By preserving fine details and spatial structure through techniques like bilinear interpolation and max pooling, PANet enhances accuracy, particularly in detecting concealed objects.

Head: The head part of YOLOv5 is same as that of YOLOv3 and YOLOv4. It consists of three convolutional layers that predict the bounding box locations, scores and object classes.

3.2 Depth-Wise Separable Convolution

A Depth-wise separable convolution is made up of two types of convolution operations. They are:

- Depth-wise Convolution,
- Point-wise Convolution.

The benefits of using Depth-wise separable convolutions are that they have lesser number of parameters to adjust as compared to the standard CNNs, which reduces overfitting. Also they are computationally cheaper as compared to the normal convolutional layers, which can help the model to be simpler and faster. The famous networks like XceptionNet [1] and MobileNet [2] used Depth-wise separable convolution.

3.2.1 Depth-Wise Convolution

A Depth-wise convolution is a convolution along only one spatial dimension of the image, whereas a normal convolution is applied across all spatial dimensions of the image. Here, since one convolutional filter is applied to each input channel, the output image will have the same number of channels as the input image.

Here we have a $10 \times 10 \times 3$ input image and a kernel of size $3 \times 3 \times 3$. These 3 channels of kernel do a convolution operation on the 3 input channels respectively. In Figure 4, the red colored channel in the input performs normal convolution with the red-color channel of the kernel and the result in the output is also represented in red. In this method, the number of filters will be the same as the number of input channels, and hence the output image will also have the same number of channels.

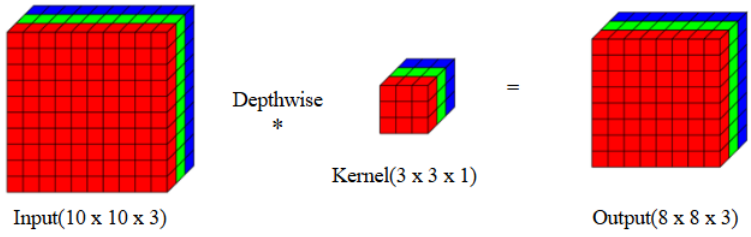


Figure 4. Depth-wise convolution operation

In Figure 4, the input image is of size $10 \times 10 \times 3$ and kernel is of size $3 \times 3 \times 3$. Since the number of input channels are 3, there are three $3 \times 3 \times 3$ kernels. The first channel of the kernel does the convolution operation with the first input channel giving one output channel, and so on. So, in the depth-wise convolution, the number of input and output channels is same.

Here, an input image of $D_f \times D_f \times M$ after applying the Depth-wise operation using kernels of size $D_k \times D_k \times M$ results in an output of size $D_p \times D_p \times N$. Each channel of the input layer is applied a normal convolution operation using a channel of kernel of size $D_k \times D_k \times 1$ and results in a channel of output of size $D_p \times D_p \times 1$. The value of D_p is

$$D_p = D_f - D_k + 1, \tag{1}$$

Input Image Size	Kernel	Output Image Size
$D_f \times D_f \times M$	$D_k \times D_k \times M$	$D_p \times D_p \times N$
$10 \times 10 \times 3$	$3 \times 3 \times 3$	$8 \times 8 \times 1$

Table 1. Output image size using Depth-wise convolution operation

where D_f is input dimension and D_k is dimension of filter.

3.2.2 Point-Wise Convolution

In a Depth-wise separable convolution, point-wise convolution is applied after the Depth-wise convolution. A point-wise convolution has a $1 \times 1 \times M$ kernel, where M is a number of input channels. See Figure 5. In the image, the output of Depth-wise convolution layer with a size $8 \times 8 \times 3$ and a kernel of size 1×1 is giving an output of $8 \times 8 \times 1$.

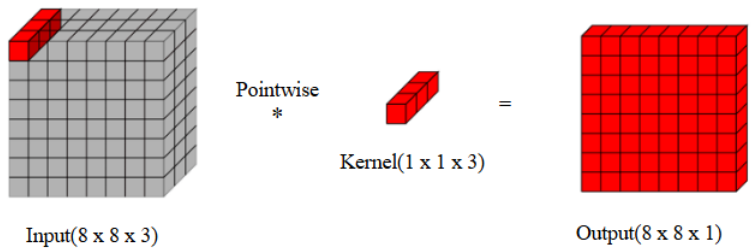


Figure 5. Point-wise convolution operation

Input Image Size	Kernel Size	Output Image Size
$D_p \times D_p \times M$	$1 \times 1 \times M$	$D_p \times D_p \times 1$

Table 2. Output image size after convolution with kernel

As shown in Table 2, the output of the Depth-wise layer is taken as input to the point-wise convolution, the input image will have a size $D_p \times D_p \times N$. The kernel of size $1 \times 1 \times M$ is used to multiply with each element in each of the $1 \times 1 \times M$ blocks of the input, as shown in Figure 5, to get the elements of the output, respectively. Here, the output image will have the same size as that of the input image of point-wise convolution and 1 channel.

A Depth-wise separable convolution has fewer parameters and fewer multiplication and addition operations as compared to a normal convolution. So Depth-wise separable convolution runs faster than a normal convolution.

The number of multiplication operations in a normal convolution layer and the Depth-wise separable convolution layer is given in Table 3. Here, we consider an

input of size $10 \times 10 \times 3$ and 1 kernel of size $3 \times 3 \times 3$, the output size is $8 \times 8 \times 1$. So, $D_f = 10$, $D_k = 3$, $M = 3$, $D_p = 8$ and $N = 1$.

	Number of Multiplications in 1 Conv Operation	Total Number of Multiplications
Normal Convolution	$D_k^2 \times M$ $3^2 \times 3 = 27$	$N \times D_p^2 \times D_k^2 \times M$ $1 \times 8^2 \times 3^2 \times 3 = 1\,728$
Depth-Wise Convolution	D_k^2 $3^2 = 9$	$N \times D_p^2 \times D_k^2$ $1 \times 8^2 \times 3^2 = 576$
Point-Wise Convolution	M 3	$N \times D_p^2 \times M$ $1 \times 8^2 \times 3 = 192$
Depth-Wise Separable (Depth-Wise + Point-Wise) Convolution	$D_k^2 + M$ $100 + 3 = 103$	$N \times D_p^2 \times (D_k^2 + M)$ $576 + 192 = 768$

Table 3. Number of multiplications in different convolutions

Here, for the given input and output size, the number of multiplications in a normal convolution is 1 728, whereas in the case of depth-wise separable convolution, the number of multiplications are 768 which is significantly less as compared to the normal convolution. There is a 55 % decrease in the number of multiplications.

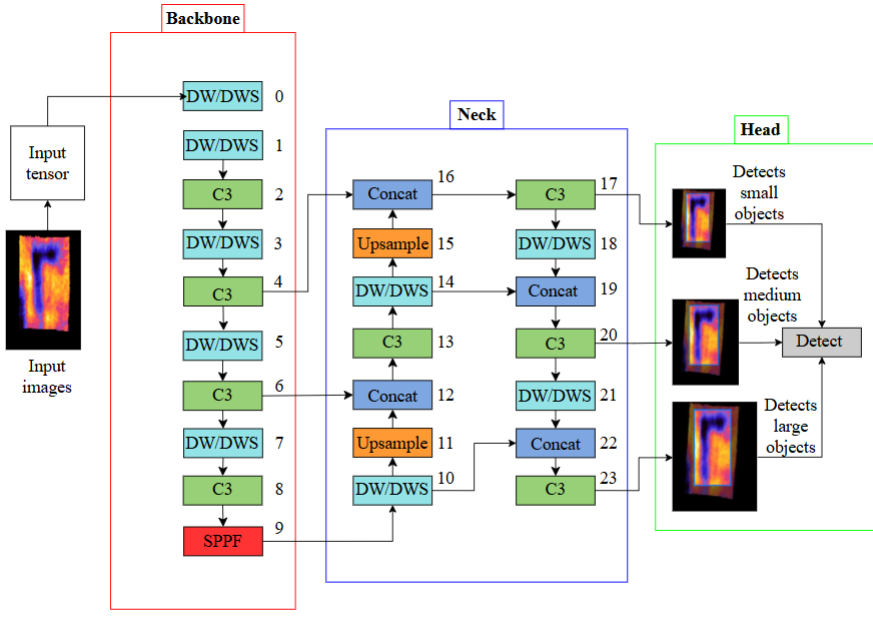


Figure 6. DW and DWS YOLOv5m architecture

Depth-wise YOLOv5 is made by replacing the Conv block in the model with DWConv, as shown in Figure 6. The DWConv consists of a Conv2d layer with groups as the greatest common divisor of input channels and output channels of the layer, BatchNormalization and SiLU activation functions.

Depth-wise Separable YOLOv5 is a custom layer made by introducing depth-wise and point-wise convolutions in place of Conv2d in Conv block. The new block contains a depth-wise convolution layer, point-wise convolution layer, BatchNormalization and SiLU activation functions.

In modified YOLOv5m architecture, the Conv blocks are replaced by DWConv blocks in case of DW-YOLOv5m and DWSConv in case of DWS-YOLOv5m.

3.3 DW-YOLOv8 and DWS-YOLOv8 Models

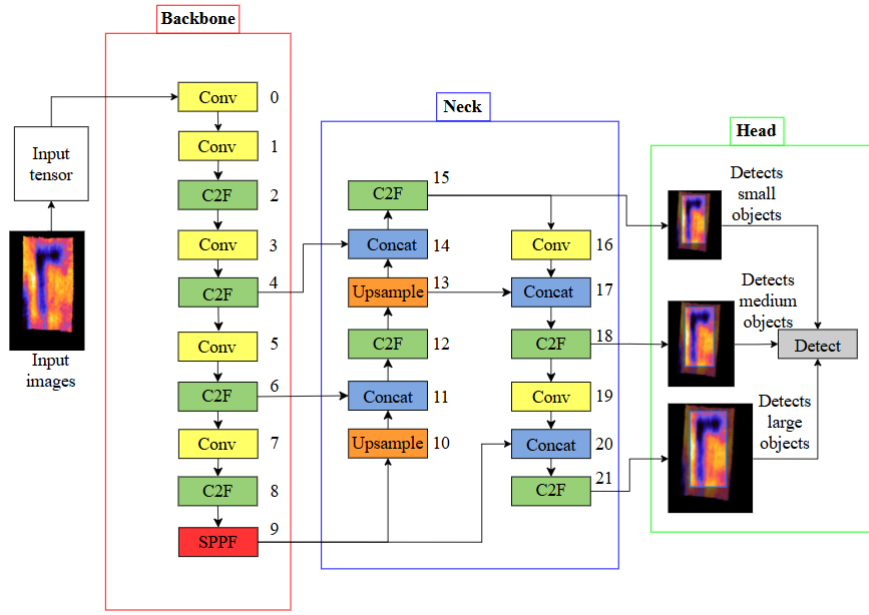


Figure 7. YOLOv8m architecture

YOLOv8 developed by Ultralytics, builds upon existing YOLO models. It outperforms earlier versions through enhancements such as spatial attention, feature fusion and context aggregation modules. These improvements lead to faster and more accurate object detection, making YOLOv8 superior to its predecessors. Key features of YOLOv8 include its improved accuracy and faster inference speed.

The architecture of YOLOv8 is shown in Figure 7. The YOLOv8 model mainly has *n, s, m, l* and *xl* variants. Each of these variants have different depth_multiple, width_multiple and max_channels. The depth_multiple determines the number of

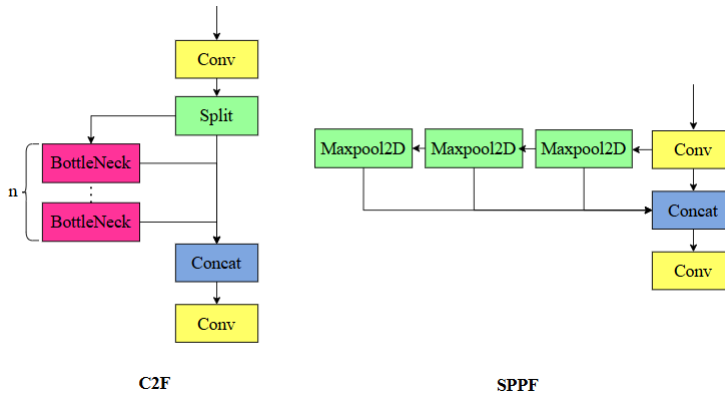


Figure 8. C2F and SPPF structure

bottlenecks in the C2F block, while the width_multiple and max_channels parameters define the output channels. These variants differ in speed and accuracy: the YOLOv8n model with its lower complexity and fewer number of parameters, offer higher speed but less accuracy. In contrast, larger models like YOLOv8l and YOLOv8xl are more complex, slower to train, and require more computational resources but provide more accurate results.

In YOLOv8, we have Conv blocks, C2F blocks, Concat, SPPF, Upsample and Detect blocks. the detect block is same as that of YOLOv5 and detects objects of various sizes. The architecture shown here corresponds to the YOLOv8m model. The YOLOv8 model is made up of the following components:

Backbone: Backbone of YOLOv8 handles the feature extraction, consisting of several convolution layers that extract the features at various levels. The C2F block, which contains multiple bottleneck layers based on the model variant (s, l, m, etc.). The layers in C2F block are conv, bottleneck and concat as given in Figure 8. The Special Pyramid Pooling Fast (SPPF) block has the same structure as that of the SPPF block in YOLOv5 architecture. It is present at the end of backbone. The main feature of SPPF is to generate representation of objects of various sizes in an image without resizing or causing spatial information loss in the image.

Neck: Neck of YOLOv8 combines the features obtained from various layers of the backbone model. At the beginning of the neck, the upsample layers increase the feature map resolution of the previous layer to match the feature map of the C2F block, to which the output of upsample layer is concatenated. The Conv block contains Conv2d, BatchNormalization and SiLU activation function.

Head: Head predicts the classes and bounding box regions which is the final output of the object detection model. In this layer, detection is performed in from three different sized feature maps to extract small, medium and large scale features

from the image. The bounding box predictions are done in YOLOv8 in the same way as that of YOLOv5.

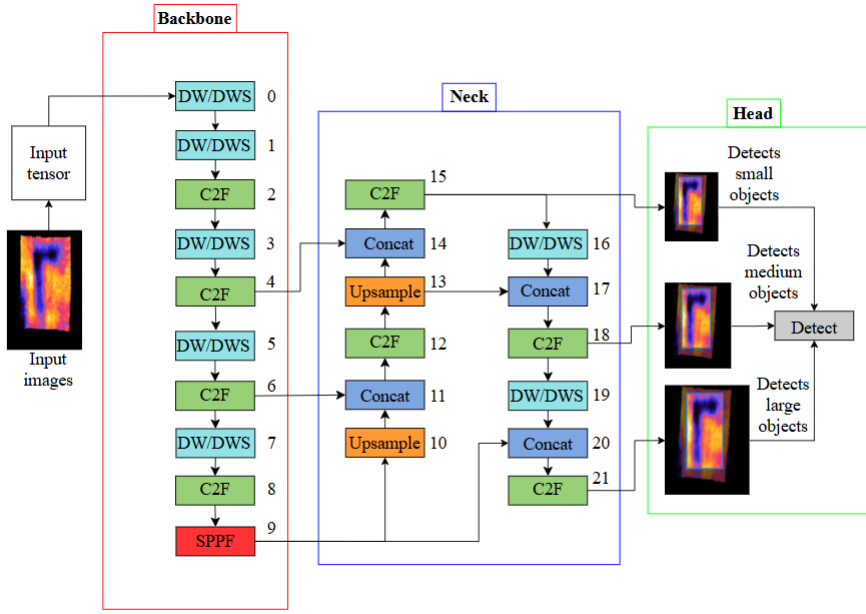


Figure 9. DW and DWS YOLOv8m architecture

In DW-YOLOv8m, the Conv blocks are replaced by DWConv blocks whereas in DWS-YOLOv8m, the Conv blocks are replaced by the custom layer DWSConv.

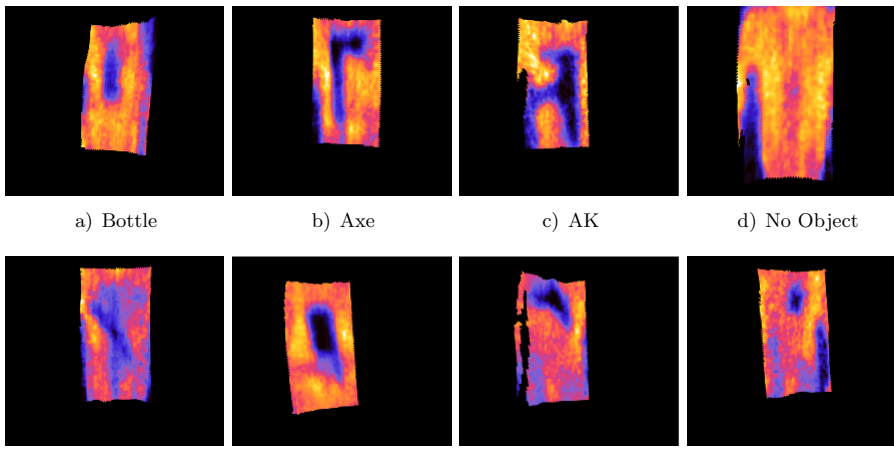
In depth-wise YOLOv8, the Conv block is replaced with a DWConv. Like the Conv block, the DWConv block includes a Conv2d layer, BatchNormalization and SiLU activation functions. However, in the DWConv block, the Conv2d layer uses groups set to the greatest common divisor of input and output channels replacing the normal Conv layer. This modification reduces the number of model parameters, simplifying the model. The modified version is shown in Figure 9.

The depth-wise Separable YOLOv5 is a custom block created by replacing Conv2d in the Conv block with depth-wise and point-wise convolution. The DWS-Conv block comprises a depth-wise separable convolution layer (depth-wise convolution layer + point-wise convolution layer), BatchNormalization and SiLU activation functions.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Dataset Description

A terahertz video dataset is used for concealed object detection and analysis [20]. Some of the images from this dataset are shown in Figure 10.



a) Bottle b) Axe c) AK d) No Object

e) Knife f) Meat Knife g) Pistol h) Cigarette Box

Figure 10. Dataset sample

From this dataset, the images are extracted by using Multimedia_04_3D_normalizer, from the diverse set of 32 videos having objects like pistol, grenade, knife and M16, etc. By consolidating videos with the same object in various parts of body into a unified class to ensure a coherent representation, the number of classes in the dataset are reduced to 22. Data augmentation is also performed to increase the size of the dataset.

These images are then annotated and processed to meet the requirements for YOLO models in Roboflow [21]. Roboflow provides a platform for annotating images by clicking and dragging to draw bounding boxes. It supports various dataset formats and allows for data augmentations, resizing, etc. After annotating and augmenting the images, the dataset can be exported in various formats.

The dataset consists of 7200 images across 22 different classes, including dangerous objects like knife, pistol, hand grenade which pose a threat when carried in public, as well non-dangerous objects like A4paper and USBDisk, as shown in Table 4. The images are divided into a training set with 5 594 images, a validation set with 794 images, and a test set with 812 images. After augmentations the training dataset has 11 188 images. The augmentation parameters are shown in Table 5

Class	Object Type	Description
1	A4paper	paper
2	AK	gun
3	AK_noMagazine	gun
4	M16	gun
5	Tin	container
6	Axe	–
7	Beltholster	pistol
8	Bottle	container
9	CandyboxLid	–
10	CigaretteBox	–
11	Fomka	–
12	Glassjar	container
13	HammerAndSickle	–
14	HandGranade	granade
15	Knife	–
16	MeatKnife	–
17	PhoneNokia	phone
18	PhoneXiaomi	phone
19	Pistol	pistol
20	SaucepanLid	–
21	Shoulderholster	pistol
22	USBDisk	–

Table 4. Object classes in the terahertz video dataset

Transformation	Range
Flip	Horizontal and Vertical
Rotation	Between $-45°$ and $+45°$
Brightness	Between $-15\,\%$ and $+15\,\%$

Table 5. Data augmentation parameters used in the experiment

4.2 Performance Metrices

Various performance matrices used for the YOLO model evaluation are:

Precision: The precision of a model is defined as the number of true positives divided by the sum of true positives plus false positives. It measures the accuracy of positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

where TP is True Positive and FP is False Positive.

Recall: The recall measures the ability of a model to correctly identify all relevant instances of a class. It quantifies the proportion of true positive predictions out

of all actual positive instances in the dataset, indicating how well the model captures the positive cases without missing any.

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

where TP is True Positive and FN is False Negative.

Mean Average Precision (mAP): The mAP metric considers both the precision and recall of the model and is calculated as the mean of the average precision for each class. A higher mAP value indicates better performance of the model.

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{4}$$

where N is the number of classes, AP_i is the average precision of class i. YOLO evaluation metrics include mAP50 and mAP50-95. mAP50 measures mean average precision at an Intersection of Union (IoU) threshold of 0.5 whereas mAP50-95 measures the mean average precision across IoU thresholds ranging from 0.5 to 0.95.

4.3 Modified YOLOv5

We trained our terahertz dataset using three models: YOLOv5m, DW-YOLOv5m and DWS-YOLOv5m. After 25 epochs, the DWS-YOLOv5m achieved the highest mAP50-95 score at 72.7 %, followed closely by DW-YOLOv5m at 72.3 %. The original YOLOv5m model attained a score of 72 %. Additionally, DW-YOLOv5 and DWS-YOLOv5m demonstrated a precision that was 0.1 % higher that the standard YOLOv5m, although their recall values were 0.1 % less than that of the original model, as shown in Table 6.

The number of parameters in DW-YOLOv5m and DWS-YOLOv5m are reduced by 26.4 % and 21.9 % respectively, compared to the original model. Due to potential fluctuations in Internet speed, affecting training time, we are not considering training time in our analysis.

Model	Precision	Recall	mAP50	mAP50-95	Layers	Trainable Parameters
YOLOv5m	0.992	1	0.995	0.72	291	20 956 179
DW-YOLOv5m	0.993	0.999	0.995	0.723	291	15 419 379
DWS-YOLOv5m	0.993	0.999	0.995	0.727	301	16 356 063

Table 6. Evaluation parameters of different YOLOv5 models on terahertz dataset

The experiments reveal that the precision, recall and mAP of the new models DW-YOLOv5m and DWS-YOLOv5m are almost same as that of the original model and the number of parameters are less than that of the original model. This shows

that the proposed models are less complex and give almost the same level of accuracy. These reductions make the models suitable for devices with lower computationally power.

4.4 Modified YOLOv8

Our dataset was trained on three different YOLOv8m models: the original YOLOv8m provided by Ultralytics and two different versions, DW-YOLOv8m and DWS-YOLOv8m. After training each model for 30 epochs, the recall value of DWS-YOLOv8m was 1 % higher than that of both YOLOv8m and DW-YOLOv8m. However, the mAP50-95 of the DWS-YOLOv5m was 1 % lower than that of DW-YOLOv5m and the original YOLOv8m model, as shown in Table 7.

The DW-YOLOv8m has the fewest parameters, with a 15 % reduction compared to the original model, followed by the DWS-YOLOv8m, which has a 17 % reduction in parameters.

Model	Precision	Recall	mAP50	mAP50-95	Layers	Trainable Parameters
YOLOv8m	0.994	0.999	0.995	0.766	295	25 869 058
DW-YOLOv8m	0.994	0.999	0.995	0.766	295	21 369 346
DWS-YOLOv8m	0.994	1	0.995	0.765	303	21 861 421

Table 7. Evaluation parameters of different YOLOv8 models on terahertz dataset

Unlike the YOLOV5m model, there is no significant improvement in mean average precision with these models. However, the recall of both DW-YOLOv8m and DWS-YOLOv8m has increased, with precision remaining unchanged. Similarly, the matrices values show no significant change. The number of parameters in both DW-YOLOv8m and DWS-YOLOv8m models are reduced compared to the original model, making them less computationally complex.

5 CONCLUSIONS AND FUTURE SCOPE

Modified YOLOv5 and YOLOv8 models are proposed in this work. Depth-wise and depth-wise separable convolutions are used to modify the models. These modified models have yielded better results compared to the original models. Despite minor variations, precision, recall, and mAP remain consistent, while the modifications significantly reduce training parameters, thereby simplifying the training process. The number of parameters is reduced by approximately 26.4 % and 21.9 % in DW-YOLOv5m and DWS-YOLOv5m respectively, compared to the original YOLOv5m model. Similarly, the reduction in the number of parameters is around 15 % and 17 % in DW-YOLOv8m and DWS-YOLOv8m respectively when compared to original YOLOv8m model. The significant reduction in the case of YOLOv5m is primarily due to the difference in the number of convolutional blocks in the model. The

lighter versions of YOLOv5 and YOLOv8, developed using Depth-wise convolution and Depth-wise Separable convolution, offer slightly better results than the original models and can operate on devices with lower computational complexity.

The proposed models have certain limitations, such as limited data, poor data quality and limited computational resources. Overcoming these limitations by developing a comprehensive terahertz dataset with diverse images and classes will significantly enhance the accuracy of concealed object detection models. As a future scope, incorporating newer models such as EfficientDet, Faster R-CNN, RetinaNet, YOLOv11 and YOLO-NAS could potentially enhance model performance.

REFERENCES

[1] Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.

[2] Howard, A. G.—Zhu, M.—Chen, B.—Kalenichenko, D.—Wang, W.—Weyand, T.—Andreetto, M.—Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. CoRR, 2017, doi: 10.48550/arXiv.1704.04861.

[3] Zeng, Z.—Luo, S.—Chen, M.—Zhao, G.—He, C.—Wu, H.: Concealed Hazardous Object Instance Segmentation for Terahertz Security Inspection Images with Channel-Spatialized Transformer. IEEE Sensors Journal, Vol. 24, 2024, No. 21, pp. 35101–35112, doi: 10.1109/JSEN.2024.3457681.

[4] Zeng, Z.—Wu, H.—Chen, M.—Luo, C. S.—He, C.: Concealed Hazardous Object Detection for Terahertz Images with Cross-Feature Fusion Transformer. Optics and Lasers in Engineering, Vol. 182, 2024, No. 3, Art. No. 108454, doi: 10.1016/j.optlaseng.2024.108454.

[5] Cheng, R.—Lucyszyn, S.: Few-Shot Concealed Object Detection in Sub-THz Security Images Using Improved Pseudo-Annotations. Scientific Reports, Vol. 14, 2024, No. 1, Art. No. 3150, doi: 10.1038/s41598-024-53045-9.

[6] Cheng, A.—Wu, S.—Liu, X.—Lu, H.: Enhancing Concealed Object Detection in Active THz Security Images with Adaptation-YOLO. Scientific Reports, Vol. 15, 2025, No. 1, Art. No. 2735, doi: 10.1038/s41598-024-81054-1.

[7] Yang, X.—Wu, T.—Zhang, L.—Yang, D.—Wang, N.—Song, B.—Gao, X.: CNN with Spatio-Temporal Information for Fast Suspicious Object Detection and Recognition in THz Security Images. Signal Processing, Vol. 160, 2019, pp. 202–214, doi: 10.1016/j.sigpro.2019.02.029.

[8] Wang, C.—Shi, J.—Zhou, Z.—Li, L.—Zhou, Y.—Yang, X.: Concealed Object Detection for Millimeter-Wave Images with Normalized Accumulation Map. IEEE Sensors Journal, Vol. 21, 2021, No. 5, pp. 6468–6475, doi: 10.1109/JSEN.2020.3040354.

[9] Wang, X.—Gou, S.—Li, J.—Zhao, Y.—Liu, Z.—Jiao, C.—Mao, S.: Self-Paced Feature Attention Fusion Network for Concealed Object Detection in

Millimeter-Wave Image. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 32, 2022, No. 1, pp. 224–239, doi: 10.1109/TCSVT.2021.3058246.

[10] Kowalski, M.: Hidden Object Detection and Recognition in Passive Terahertz and Mid-Wavelength Infrared. Journal of Infrared, Millimeter, and Terahertz Waves, Vol. 40, 2019, No. 11, pp. 1074–1091, doi: 10.1007/s10762-019-00628-7.

[11] Pang, L.—Liu, H.—Chen, Y.—Miao, J.: Real-Time Concealed Object Detection from Passive Millimeter Wave Images Based on the YOLOv3 Algorithm. Sensors, Vol. 20, 2020, No. 6, Art. No. 1678, doi: 10.3390/s20061678.

[12] Danso, S. A.—Shang, L.—Hu, D.—Odoom, J.—Liu, Q.—Nyarko, B. N. E.: Hidden Dangerous Object Recognition in Terahertz Images Using Deep Learning Methods. Applied Sciences, Vol. 12, 2022, No. 15, Art. No. 7354, doi: 10.3390/app12157354.

[13] Jayachitra, J.—Devi, K. S.—Manisekaran, S. V.—Satti, S. K.: Terahertz Video-Based Hidden Object Detection Using YOLOv5m and Mutation-Enabled Salp Swarm Algorithm for Enhanced Accuracy and Faster Recognition. The Journal of Supercomputing, Vol. 80, 2024, No. 6, pp. 8357–8382, doi: 10.1007/s11227-023-05717-y.

[14] Xu, F.—Huang, X.—Wu, Q.—Zhang, X.—Shang, Z.—Zhang, Y.: YOLO-MSFG: Toward Real-Time Detection of Concealed Objects in Passive Terahertz Images. IEEE Sensors Journal, Vol. 22, 2022, No. 1, pp. 520–534, doi: 10.1109/JSEN.2021.3127686.

[15] Ge, Z.—Zhang, Y.—Wu, X.—Jia, Z.—Wang, H.—Jia, K.: Deep-Learning-Based Method for Concealed Object Detection in Terahertz (THz) Images. Advanced Fiber Laser Conference (AFL2023), Proceedings of SPIE, Vol. 13104, 2024, doi: 10.1117/12.3021687.

[16] Wang, T.—Gao, S.—Huo, Y.—Cattani, P.—Mei, S.: Depthwise Separable Axial Asymmetric Wavelet Convolutional Neural Networks. Vol. 163, 2024, doi: 10.1016/j.asoc.2024.111886.

[17] Panigrahi, S.—Raju, U. S. N.: DSM-IDM-YOLO: Depth-Wise Separable Module and Inception Depth-Wise Module Based YOLO for Pedestrian Detection. International Journal on Artificial Intelligence Tools, Vol. 32, 2023, No. 04, Art. No. 2350011, doi: 10.1142/S0218213023500112.

[18] Qin, Y. Y.—Cao, J. T.—Ji, X. F.: Fire Detection Method Based on Depthwise Separable Convolution and YOLOv3. International Journal of Automation and Computing, Vol. 18, 2021, No. 2, pp. 300–310, doi: 10.1007/s11633-020-1269-5.

[19] Liu, T.—Pang, B.—Zhang, L.—Yang, W.—Sun, X.: Sea Surface Object Detection Algorithm Based on YOLO V4 Fused with Reverse Depthwise Separable Convolution (RDSC) for USV. Journal of Marine Science and Engineering, Vol. 9, 2021, No. 7, Art. No. 753, doi: 10.3390/jmse9070753.

[20] Morozov, A. A.—Sushkova, O. S.: Development of a Publicly Available Terahertz Video Dataset and a Software Platform for Experimenting with the Intelligent Terahertz Visual Surveillance. In: Bhattacharjee, D., Kole, D. K., Dey, N., Basu, S., Plewczynski, D. (Eds.): Proceedings of International Conference on Frontiers in Computing and Systems (COMSYS 2020). Springer, Singapore, Advances in Intelligent

Systems and Computing, Vol. 1255, 2021, pp. 105–113, doi: 10.1007/978-981-15-7834-2_10.

[21] Dwyer, B.—Nelson, J. et al.: Roboflow (version 1.0) [software]. 2022, `https://roboflow.com`.

Singara Singh Kasana is Professor in the Department of Computer Science and Information Technology, Central University of Haryana, Mahendergarh, India. Before joining the Central University of Haryana, he worked at the Thapar Institute of Engineering and Technology, Patiala, India, for nearly 19 years. With over 24 years of teaching and research experience, he earned his Ph.D. degree in image compression from Thapar University. He has supervised 8 Ph.D. theses and 35 Post Graduate dissertations. His research interests include digital twins, digital image processing, machine learning, image and video forensic, and computer vision. He has published more than 70 research papers in reputed international journals and conferences.

Lakshmy Santhosh is a Software Engineer at the Tata Consultancy Services. She completed her M.Sc. in data science from the Central University of Haryana and her B.Sc. in physics from the Miranda House, University of Delhi. Her areas of interest include machine learning, artificial intelligence, deep learning, and image processing.