**International Journal of Electrical and Computer Engineering (IJECE)**
Vol. 13, No. 3, June 2023, pp. 2812~2826
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i3.pp2812-2826

# Pedestrian classification on transfer learning based deep convolutional neural network for partial occlusion handling

**May Thu[1,2], Nikom Suvonvorn[3], Nichnan Kittiphattanabawon[1]**
[1]School of Informatics, Walailak University, Nakhon Si Thammarat, Thailand
[2]Informatics Innovative Center of Excellence, Walailak University, Nakhon Si Thammarat, Thailand
[3]Department of Computer Engineering, Prince of Songkla University, Songkhla, Thailand

**Article Info**

***Article history:***

Received Jul 12, 2022
Revised Sep 12, 2022
Accepted Oct 1, 2022

***Keywords:***

Convolutional neural network
Partial occlusion handling
Pedestrian
Pedestrian dataset
Transfer learning

## ABSTRACT

The investigation of a deep neural network for pedestrian classification using transfer learning methods is proposed in this study. The development of deep convolutional neural networks has significantly improved the autonomous driver assistance system for pedestrian classification. However, the presence of partially occluded parts and the appearance variation under complex scenes are still robust to challenge in the pedestrian detection system. To address this problem, we proposed six transfer learning models: end-to-end convolutional neural network (CNN) model, scratch-trained residual network (ResNet50) model, and four transfer learning models: visual geometry group 16 (VGG16), GoogLeNet (InceptionV3), ResNet50, and MobileNet. The performance of the pedestrian classification was evaluated using four publicly datasets: *Institut National de Recherche en Sciences et Technologies du Numérique* (INRIA), Prince of Songkla University (PSU), CVC05, and Walailak University (WU) datasets. The experimental results show that six transfer learning models achieve classification accuracy of 65.2% (end-to-end CNN), 92.92% (scratch-trained ResNet50), 97.15% (pre-trained VGG16), 94.39% (pre-trained InceptionV3), 90.43% (pre-trained ResNet50), and 98.69% (pre-trained MobileNet) using data from Southern Thailand (PSU dataset). Further analysis reveals that the deeper the ConvNet architecture, the more specific information of features is provided. In addition, the deep ConvNet architecture can distinguish pedestrian occluded patterns while being trained with partially occluded parts of data samples.

***Corresponding Author:***

Nichnan Kittiphattanabawon
School of Informatics, Walailak University
Tha Sala, Nakhon Si Thammarat, 80160, Thailand
Email: knichcha@wu.ac.th

## 1. INTRODUCTION

Annually, over 1.3 million people, have died in traffic accidents around the world, with collision rates surging in most developed countries [1]–[3]. In Thailand, the coronavirus disease (COVID-19) pandemic is not the country's most serious public health issue, according to the number of deaths on the road. As of mid-September, the COVID-19 pandemic had claimed 60 deaths since it hit the country in 2020. The number of COVID-19 deaths in the last six months is comparable to the number of traffic fatalities in a single day [2], [3]. Also, Thailand's road accidents had the second-highest rate, and most of the dead were pedestrians and motorcyclists. The World Health Organization (WHO) states that Thailand's road accidents are among the worst in Southeast Asia. Approximately 20,000 people are killed in traffic accidents every year, equaling about 56 deaths per day. Pedestrian classification has been a trend in computer vision research topics in the past decades, focusing on classifying, detecting, tracking, and analyzing objects such as pedestrians or vehicles. Additionally, there are a variety of applications, including video surveillance, intelligent transportation systems, and self-driving assistance systems. Though there are many significant improvements for detecting pedestrians, the crucial yet challenging problems include the large variety of poses, appearances, sizes, and types of occlusions [4]–[6]. Among these problems, partial occlusion frequently occurs due to the diversity of the partially occluded patterns of the pedestrians, and it needs further investigation between the pedestrians and the crowded instances.

Handcrafted feature representations were previously used by the following traditional pedestrian detectors: Haar [7], [8], scale invariant feature transform (SIFT) [9], [10], histogram of oriented gradient (HOG) [11]–[13], and local binary pattern (LBP) [7] [14]. To perform pedestrian classification, these feature representations are combined with classifiers such as support vector machine (SVM) [15], [16] boosted forests [10], and AdaBoost [17]. However, detection difficulties arise when the pedestrian and the attached instances are partially obscured. The advanced deep convolutional neural networks (deep ConvNets) have captivated a lot of attention in automatic hierarchical learning with effective rich representations to expand the ability of handcrafted features. Due to the hierarchical representation of discriminating features retrieved by multi-layered neural networks, deep ConvNets have successfully conducted the classification of the object in machine vision compared to handwritten features. Convolutional neural network (CNN) architecture is a type of deep neural network commonly used to learn images and videos. CNN comprises the following layers: an input layer, a pooling layer, a fully connected layer, and an output layer that learns deep patterns and structures from example input. CNN has achieved continuous improvement in image recognition and classification through the establishment of innovative layers for visual geometry group 16 (VGG16), VGG19, GoogLeNet (Inception-v1, v2, v3, v4), residual network (ResNet50), and Inception-ResNets, MobileNet architectures, respectively [18]–[22]. Primarily, the ImageNet challenges on various datasets are used deep CNN. Several deep CNN models have gained much attention for pedestrian detection and achieved extreme performances in the occlusion handling tasks [23]–[26]. Furthermore, on a large-scale ImageNet dataset, the deep CNN models were extensively tweaked for image classification and obtained great classification performance. However, the performance of the deep CNN can influence the results on a small amount of training data and the presence of partially occluded pedestrian areas in the training data [4].

Deep convolutional neural networks require hundreds of thousands of sample images, and the number of samples has an effect on learning performance. Transfer learning in deep learning is an approach whereby the training time expenses of the deep CNN model is being solved, and a small number of training samples is also solved for the specific complex models [4], [24], [27]. Furthermore, transfer learning minimizes the time required to train a neural network model and can lead to fewer generalization errors. In addition, the transfer learning model can be used as a feature extraction process and a classification process when training a new model. The quality of the collected images influences the outcome of the classification process in pedestrian classification, particularly the variation of light, the properties of the equipment (camera, mobile camera, and sensors), the conditions of the environment, the variation of appearances, and the types of occlusions. Road accidents were common in Southern Thailand, particularly during the Songkran festivals. In a real-time setting, the pedestrian crossing sign and speed monitoring are critical for reducing the number of incidents. The diverse perspectives of pedestrians with a lot of vehicles were included in the data samples acquired for this study (cars, motorcycles, tuk-tuk, and tri-motor cycles). Pedestrians with diverse occlusions under complex backgrounds are the most difficult aspects of the classification procedure. With the effective use of the transfer learning approach, this research aims to solve the pedestrian classification of real-time data samples in Southern Thailand.

The motivation for this study is to examine the capability of the deep convolutional neural network in pedestrian classification for dealing with partially occluded parts of pedestrians. Hence, we concentrate on the following research problems in order to address pedestrians' partially occluded parts; i) how can transfer learning techniques help in improving pedestrian classification? and ii) how well do various neural networks perform in pedestrian classification models? The main contributions of this research can be summarized as follows. Firstly, Walailak University (WU) dataset is introduced for research purposes to achieve more performance on partial occlusion handling. We aim at finding the suitable architecture for applying the pedestrian detection system in a real-world environment. Secondly, this paper discusses two commonly used transfer learning strategies (feature extraction and fine-tuning). To answer the research questions, six classification models are developed and conducted for pedestrian classification, including the end-to-end CNN model, scratch-trained ResNet model, and four transfer learning models: VGG16, GoogLeNet (InceptionV3), ResNet50, and MobileNet with ImageNet weighted parameter values. Moreover, various optimizers and learning parameters are employed to examine pedestrian classification. Initially, an end-to-end CNN model with ten convolution layers, four max-pooling layers, and a fully connected layer is built. As the second model, we trained ResNet50 with two small datasets: *Institut National de Recherche en Sciences* *et Technologies du Numérique* (INRIA) and Prince of Songkla University (PSU). Following that, these networks were trained for two classes (pedestrian and non-pedestrian) that were supplied as input in images and then tested with various images. For the remaining four transfer learning models, we leveraged the pre-trained CNN models and then constructed a new customized fully connected layer for pedestrian classification. Finally, we examine and contrast the various networks with varying categorization performance. In this study, different neural network architectures have distinct layers, and their classification performance varies enormously. The classification accuracy of the deep convolutional neural network can be checked using with the complex real-world environment datasets.

## 2. RELATED WORKS

Over the past few decades, the rapid improvement of deep learning methods achieved more comprehensive performance tasks in pedestrian detection. In computer vision and image classification, the development of deep CNN aimed to use in hierarchy learning levels with the in-depth features of low-level vectors and high-level semantic parts from a fixed-size input. Compared with the handcrafted features, deep convolutional features automatically extracted features with its multi-layer hierarchical representation of data and parallel computing of end-to-end learning mechanisms. Images can be classified more accurately with the deep CNN. LeNet-5 architecture is a rudimentary convolutional neural network that uses gradient-based learning algorithms to recognize handwritten characters with minimum pre-processing. When compared to traditional pattern recognition approaches, multi-layered networks significantly outperformed with feed-forward feature extraction on complex hand-designed information. The Hebbian principle "neurons that fire together, wire together" was then inspired to create the deeper layer of the network in the form of the inception module, which increased the neural network's representation power. The inception architecture of GoogLeNet, the best-performing entry in the ImageNet large scale visual recognition challenge (ILSVRC-2014) classification competition, was built using an extremely deep convolutional neural network with 22 layers and tiny convolutional filters with 1×1 and 5×5 convolutions. Later, Microsoft proposed ResNet architecture, a residual learning framework with up to 152 convolutional layers that earned first place in the ILSVRC-2015 classification competition. It is not necessary to include any additional parameters in order to use identity shortcut connections to execute identity mapping and reduce computational complexity.

Cheng *et al*. [8] also presented a fast fused part-based model (FFPM) for producing deep spatial features using six AdaBoost classifiers and Haar-like features for distinct sections of the body. The proposed model is divided into three parts: a part-boosting model, a full-body boosting model, and an n number of classifiers. Part-based features were taken from the six body parts of pedestrians (head, left shoulder, right shoulder, hands, knees, and feet) for the six adaptive boosting models in the part-boosting model. To acquire the information of pedestrians, these collected features were fed into the full-body boosting model. The root-boosting model and the support vector machine (SVM) classification result determined the final detection result. The experiment results on the proposed method attained 83.1% improvement in pedestrian detection. Chen *et al.* [9] proposed a multi-layer fused convolution neural network (MLF-CNN) which consists of two-stream networks: a proposal network and a detection network to improve the detection performance of pedestrians under adverse illumination. To reduce the detection, miss rate, the multi-layer fusion region proposal network (RPN) was employed to match the scale utilizing fused region of interest (ROI) pooling, which extracted three feature maps instead of a single-layer extraction. The detector's classification was improved further by using the fused ROI pooling layer in a variety of environments. The proposed architecture outperformed the existing approaches in the performance accuracy with the several challenging datasets. Lin *et al*. [25] introduced PedJointNet, a two-branch network architecture for pedestrian recognition that specifies both the full body and the head-shoulder component. Two feature pyramid modules were combined to create the branches, one for the head-shoulder parts of pedestrians and the other for the entire parts of pedestrians. To improve the network across feature maps learning, feature pyramid network (FPN) and atrous spatial pooling (ASSO) are commonly used. Multi-level prediction branches were also designed depending on the specific performance of each part in the detection. Adaptive fusion layers are employed to alter the weighted loss for the final layer by concatenating the two branches.

Luo *et al*. [28] introduced the switchable deep network (SDN) with the mixture representation of several parts (i.e., body, head, shoulder, upper body, and lower body) to explicitly identify the complicated mixture of visual variations at multiple levels, which was inspired by the part-based model. It was able to interpret the most suitable templates of each component for the mixture models and the entire body at the part and body levels. The hierarchical features, salience maps, and mixture representation of the body parts were all learned by the proposed model. Wang *et al.* [29] presented the part and context network (PCN), which uses two sub-networks to recognize pedestrians using semantic and context information from body parts. The semantic parts information was completely used for precise information in the part branch, and the pedestrian box was divided into several part grids after ROI pooling, with each part grid associated with a detection score. Later, long short-term memory (LSTM) was used to communicate pedestrian segments using those semantic scores. Different context areas with varied scales were employed to extract context information in a context branch because different pedestrian instances may require different context information for the heavy-occluded. By merging the results of all branches for occlusion handling, an effective interdependent detector has been constructed.

Using the soft-rejection-based network fusion method, Du *et al*. [30] introduced the fused deep neural network (FDNN) with concurrent processing of several networks. A pedestrian candidate generator, a classification network, and a pixel-wise semantic segmentation network made up the prospective architecture design. As an object detector, a single shot deep convolutional network constructed all conceivable pedestrian candidates of various sizes and occlusion for detecting all actual pedestrians. The aggregated degree of confidence in those candidates from the classifiers improves the pedestrian candidates' confidence scores, and the classification network comprises of several binary classifiers. To obtain the final confidence scores, a soft-rejection-based network fusion method was devised to fuse the soft metrics from the networks. To improve pedestrian identification performance, Park *et al.* [31] presented the probabilistic fusion network, which comprises of three subnetworks: feature extraction network, region proposal network, and inference network. A channel weighting fusion layer (CWF) and an accumulated probability fusion layer (APF) were also proposed to train these three sub-networks into a single network at the same time. On multi-spectral pedestrian datasets, the proposed technique improved the existing approaches. The multi-scale CNN (MS-CNN) was introduced by Cai *et al*. [32] to provide an accurate object proposal on a detection network using feature up-sampling. Two sub-networks: a proposal sub-network and a detection sub-network, have been implemented on the CNN-based detector for end-to-end learning. Object detection was used at several output layers to fit the varying scales of objects and receptive fields. The output layer combination was used as complementary detectors for a strong multi-scale detector by improving a multi-task loss. In addition, a feature up-sampling was employed instead of input up-sampling at the deconvolution layer for segmentation and edge detection to conserve memory and convolutional costs. On the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) benchmark testing dataset, the proposed technique scored 83.92% for pedestrians (easy), 73.70% for pedestrians (moderate), and 68.31% for pedestrians (hard). The performance evaluation achieved high detection rates for fast multi-scale object detection. Previous deep learning-based pedestrian identification systems indicated that deep models had a dramatic impact on occlusion handling due to the hierarchical learning process's extensive usage of rich feature representations. On the other hand, the widespread use of large datasets necessitates competitive performance in comparison to other approaches as well as a long computing time to solve complex models. Deep ConvNets models solve the occlusion problems and can get better performance results while handling the moderate occluded pedestrians.

## 3. METHOD

Over the several years, robust deep network architectures have been widely used in the pedestrian detection system with their rich representation of structural learning, and many researchers have been conducted the deeper structure of the baseline architecture. There are dozens of top-performing models related to computer vision tasks for image classification. Transfer learning is defined in this context as a form of weight initialization strategy. While this may be useful in situations where the relevant problem contains less amount of labelled data, it may also be effective the performance accuracy of the context. VGG (VGG16), GoogLeNet (InceptionV3), ResNet50, and depthwise separable convolution MobileNet are some of the more prominent classification models. Their performance as well as their architectural improvements make these models a popular choice for transfer learning. The objective of our study is to analyze the performance of the networks using collected pedestrian data from real-time environment. The following section goes into great detail about each of these architectures.

### 3.1. End-to-end convolutional neural network

An end-to-end CNN model is trained on two publicly datasets: the INRIA pedestrian dataset and the PSU pedestrian dataset. The architecture of the designed CNN model is depicted in Figure 1. CNN enables automatic feature extraction throughout its layers and retains the input as raw data without performing any special standardization. There are fifteen layers in this convolutional neural network model. The first layer is the input layer with the size of 128×128. There are ten convolution layers in CNN, which are the primary layers of the architecture. The convolution operation has the role of extracting distinct features from the input, each of which contains the raw pixel values of the image. The initial convolution layers are responsible for obtaining low-level features such as edges, lines, and corners. There are numerous kernels in each convolutional layer, and each kernel is replicated across the entire image with the same parameters as the previous kernel. The 3×3 filters are used to learn different feature types in these layers. After each convolution, the rectified linear units (ReLU) layers are used to swap the negative number of the pooling layer and remove the noisy regions. To minimize the feature dimensions and overfittings, a 2×2 filter with stride 2 is applied after the convolutional layer in the average-pooling layer. The final layer of a CNN model is the fully connected layer, which is responsible for computing the class values for a given input. As an output, the softmax layer classifies the maximum probability of a given normalized class (pedestrian or non-pedestrian).

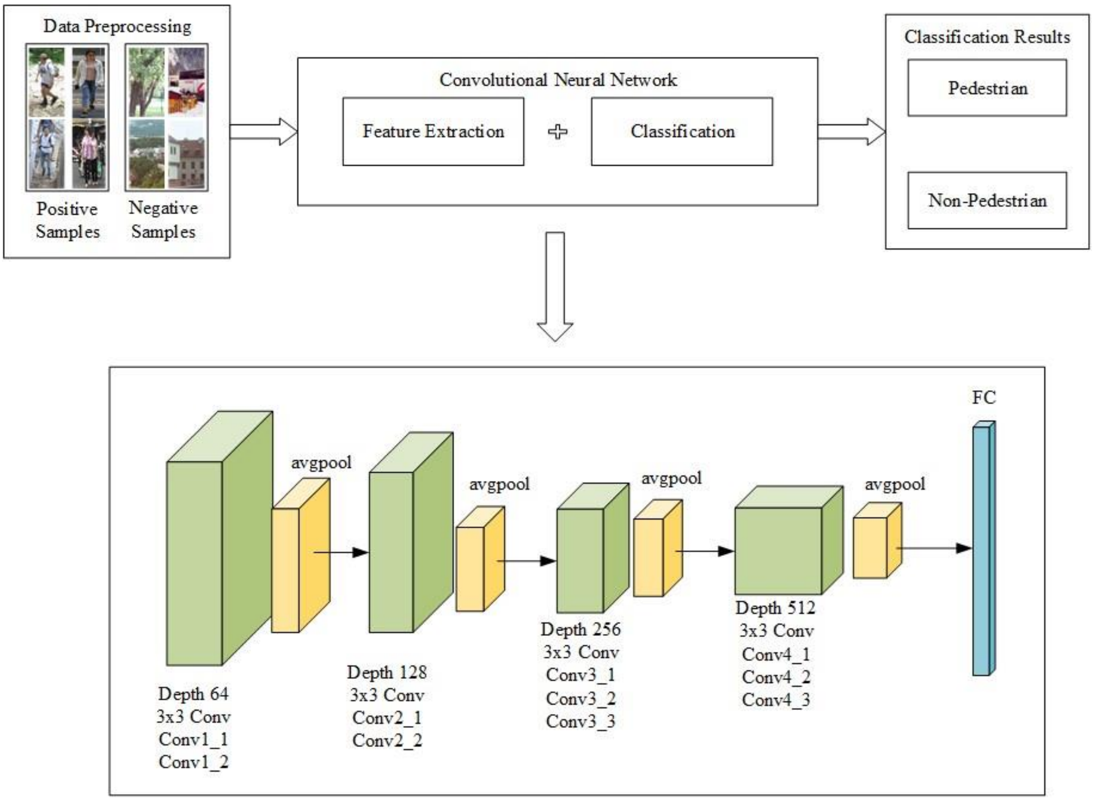


Figure 1. The architecture of the end-to-end convolutional neural network

### 3.2. Deep feature extraction model

The ResNet50 convolutional neural network model is employed in the deep feature extraction model, and it is scratch-trained on pedestrian datasets from INRIA and PSU for the pedestrian classification task. In terms of dealing with the concept of residual learning, Microsoft launched ResNet, a classic backbone network, for use in a variety of computer vision tasks [21]. In particular, ResNet50 presented a deep residual network framework for the concept of skip connections (shortcut connections and residual) as an alternative to skipping every few stack layers. There are five phases in all, including a convolution (CONV) and an identity block (ID Block). Each CONV block is comprised of three convolutional layers, and each ID block is built up of three convolutional layers as well. When the depth of the network increases, the Residual network is used to optimize the larger training error problem, which is achieved by the use of the residual block in ResNet architecture. ResNet has adapted batch normalization (Batch Norm), the CONV and ID blocks are the fundamental building elements of the ResNet architecture. Furthermore, feedforward networks with shortcut connections can be used to achieve the formation of F(x) + x by skipping one or more layers of the network. Essentially, these shortcut connections accomplished identity mapping by adding the outputs of this mapping to the outputs of the stacked layers for cantering layer responses, gradients, and propagated errors, which resulted in the identity mapping being completed. To preserve the time complexity per layer, the feature map is directly downsampled with stride convolution, and a Batch Norm is applied right after each convolution and before activation, as in the architecture. The network also includes a global average pooling (AVG POOL) layer as well as 1,000 fully connected layers with softmax at its final result. The overall architecture of the ResNet50 model is demonstrated in the Figure 2.

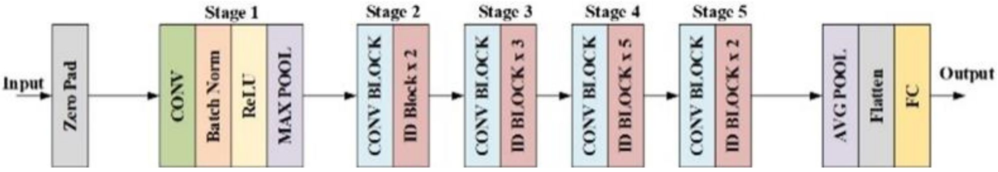


Figure 2. The architecture of the ResNet50 feature extraction model

### 3.3. Pre-trained convolutional neural network

If there is a minimal amount of training data in a problem domain, training a CNN from scratch is inefficient. The most acceptable technique for this problem is to use a pre-trained model collected from a big dataset for a problem with a specific dataset. Therefore, we apply fine-tuning for VGG 16, InceptionV3, ResNet50, and MobileNet models trained with the ImageNet image database. The architecture of these pre-trained models is shown in Figure 3.

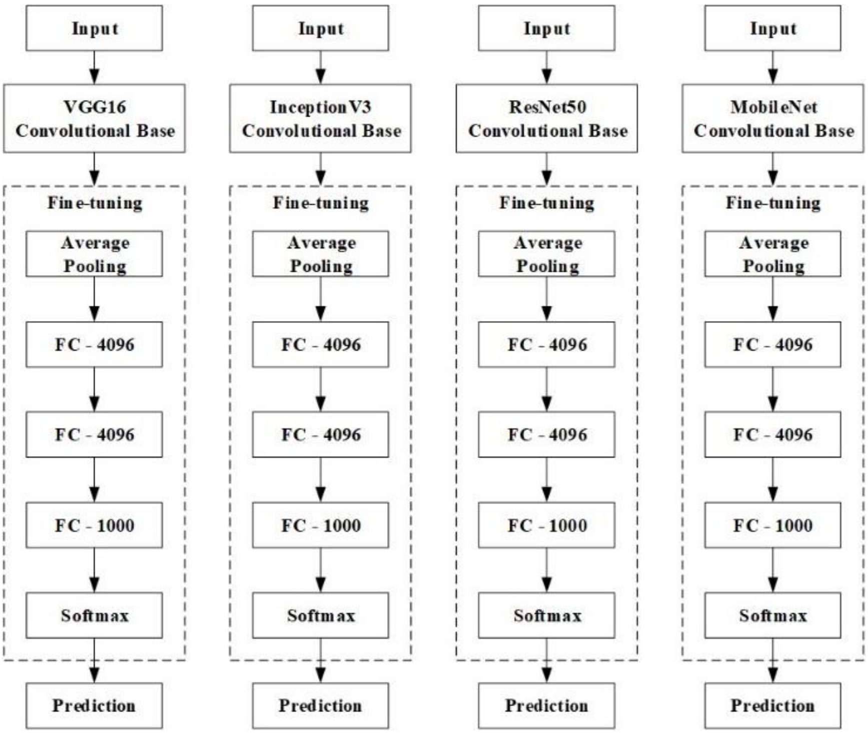


Figure 3. The architecture of the pre-trained convolutional neural network model

To fine-tune the models, we freeze the entire layer of the base model in order to harness the knowledge, and the model's weights are initialized using the trained ImageNet dataset weights. The last three layers of the deep CNN models have 1,000 classes. These three layers are therefore eliminated and replaced with those that are more appropriate for the pedestrian dataset. Following that, a new fully connected layer is created with the two outputs and the number of classes in the new training dataset. To accelerate the training phase for hyper-parameter tuning, learning rates of 0.00001 with the Adam optimizer and the rectified linear unit (ReLU) are utilized. A dropout layer with a parameter of 0.5 is added to reduce overfitting and improve the generalization of the network. In general, the pre-trained model's training procedure is as follows: i) use the pedestrian classification task to retrain the deep network with an input size of 224×224 and standard color augmentation, i.e., utilizing the image annotations of two classes from the INRIA and PSU datasets and ii) in the fully connected (FC) layer, fine-tune the network for the pedestrian classification task for data classification (two classes). Lastly, we added a flattened layer, two fully connected dense layers, and a dropout layer with a probability of 0.5 to the final FC layer. The probabilities of a given normalized class are used to determine whether the final softmax layer output is "pedestrian" or "non-pedestrian".

## 4. RESULTS AND DISCUSSION

The experiment setup and procedures for the proposed techniques, as well as the evaluation outcomes, are explained in this section. The results of the experiment were divided into three categories: end-to-end convolutional neural network models, deep feature extraction models, and pedestrian classification models that had been pre-trained. The experiment setup was conducted on Intel® Core™ i7-4770 CPU @ 3.40 GHz, NVIDIA Quadro RTX4000 GPU, 32 GHz RAM, and Windows 10pro environment using OpenCV 4.4.0 library and Keras with TensorFlow 2.0 backend. The datasets are separated into training and testing sets in proportions of 70% and 30%, respectively. During the training phase, all deep learning-based models have 100 epochs. The computational time of each deep learning-based model are given in Table 1.

Table 1. Elapsed time results of deep learning-based models

| Methods | Elapsed Time | |
|---|---|---|
| | One Epoch (second) | Step-per-Epoch (second) |
| End-to-End CNN | 2 | 0.019 |
| Scratch-trained ResNet50 | 3927 | 2 |
| Pre-trained ResNet50 | 1210 | 0.663 |
| Pre-trained MobileNet | 433 | 0.237 |
| Pre-trained InceptionV3 | 584 | 0.319 |
| Pre-trained VGG16 | 81 | 0.045 |

### 4.1. Datasets

In this experiment, two publicly available pedestrian datasets are used to conduct research. We performed our analysis on the INRIA, PSU, CVC05, and WU datasets because these four datasets contained raw images with annotations and different resolutions of image pixels with positive and negative images in the training and testing set. In the experiment, INRIA and PSU datasets were used for the training the proposed models meanwhile, CVC05 and WU datasets were used as the evaluation tests. Table 2 summarizes the characteristics of the pedestrian datasets.

Each pedestrian dataset in Table 2 has two classes, which are divided into training and testing sets, respectively. The following properties were typically collected: scene, perspective, pose, and occlusion with attached items in a marketplace, a city center intersection, and along university campus roadways. The INRIA person dataset contains high-resolution upright pedestrian photographs collected from various directions, separated into training and testing samples, with annotations provided for both positive (pedestrian) and negative images (those not containing pedestrians). All of the pictures were taken from a personal collection and depict pedestrians in a variety of poses upon a variety of surroundings (indoors, urban, rural), despite the fact that they were generally standing or walking. The INRIA data samples are illustrated in Figure 4. Examples of the positive (pedestrian) samples can be seen in Figures 4(a) and 4(b) explain the examples of the negative (non-pedestrian) samples.

Table 2. Properties of the experimented pedestrian datasets

| Datasets | Imaging Setup | Training | | | Testing | | | Properties | | | | | | Publication |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #Pedestrians | #neg. images | #pos. images | #Pedestrians | #neg. images | #pos. images | Color images | Occlusion labels | Illumination | Video seqs. | View labels | Pose labels | |
| INRIA | Photo | 1208 | 614 | 1218 | 556 | 288 | 453 | √ | | | | | | 2005 |
| CVC05 | Photo | - | - | - | 593 | 700 | 593 | √ | √ | √ | | √ | √ | 2013 |
| PSU | Photo | 1186 | 517 | 1051 | 1270 | 517 | 755 | √ | √ | √ | | √ | √ | 2018 |
| WU | Photo | 2021 | 7670 | 2021 | 1585 | 1142 | 1000 | √ | √ | √ | √ | √ | √ | 2022 |

The PSU dataset includes images with partial or serious occlusion in a variety of positions and complex backgrounds. Data was taken from pedestrians in a variety of postures, including upright, walking, standing, cycling, motorbike riding left, right, back, and occluded areas. It has four resolutions: 64×64, 256×256, 720×960, and 960×720 pixels, all downsized from a 3120×3120 pixel original. Some examples from the PSU dataset are illustrated in Figure 5. Examples of the positive (pedestrian) samples can be seen in Figures 5(a) and 5(b) explain the examples of the negative (non-pedestrian) samples.

The CVC05 partially occluded pedestrian dataset was created under partial occlusion at the per-image level. Pedestrian images were captured with a digital camera with 640×480 pixels resolution in the urban scenarios of Barcelona. Pedestrians with mirror images and non-pedestrian images were collected from the road area. The CVC ADAS group annotated both non-pedestrian photographs and partially occluded pedestrian images. Some example data samples from the CVC05 dataset are shown in Figure 6. Examples of the positive (pedestrian) samples can be seen in Figures 6(a) and 6(b) explain the examples of the negative (non-pedestrian) samples.

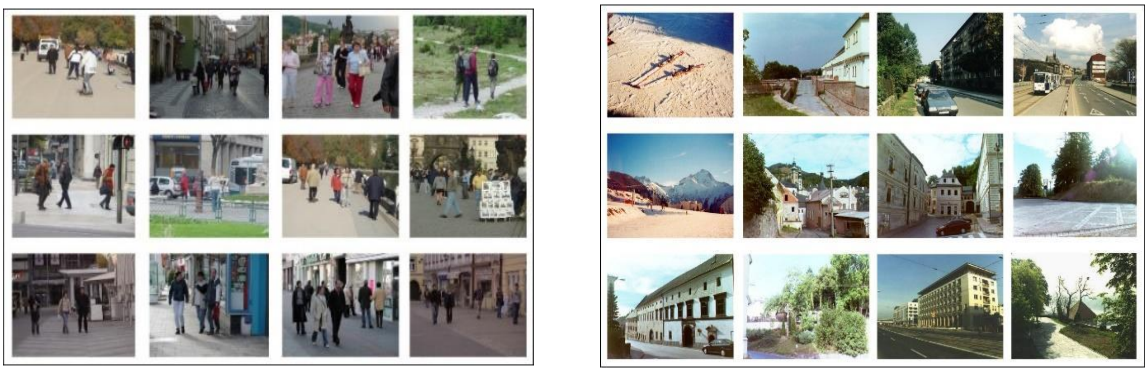

(a) (b)

Figure 4. Illustration of the INRIA dataset: (a) examples of the positive (pedestrian) samples and (b) examples of the negative (non-pedestrian) samples

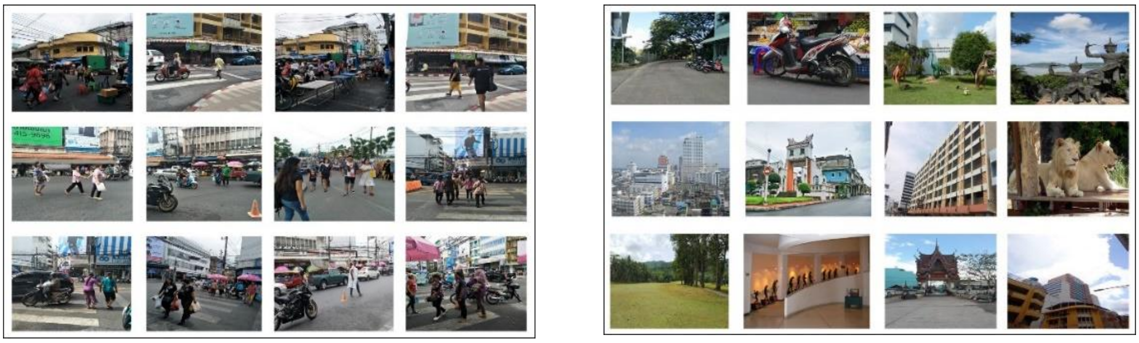

(a) (b)

Figure 5. Illustration of the PSU dataset: (a) examples of the positive (pedestrian) samples and (b) examples of the negative (non-pedestrian) samples

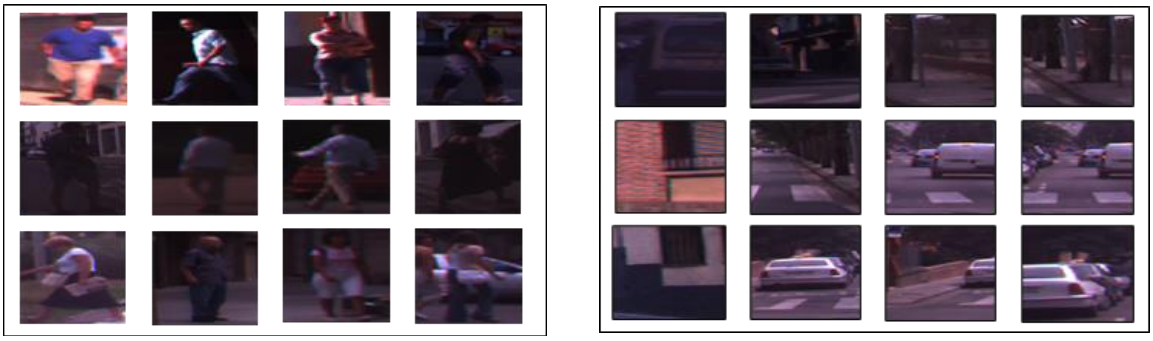
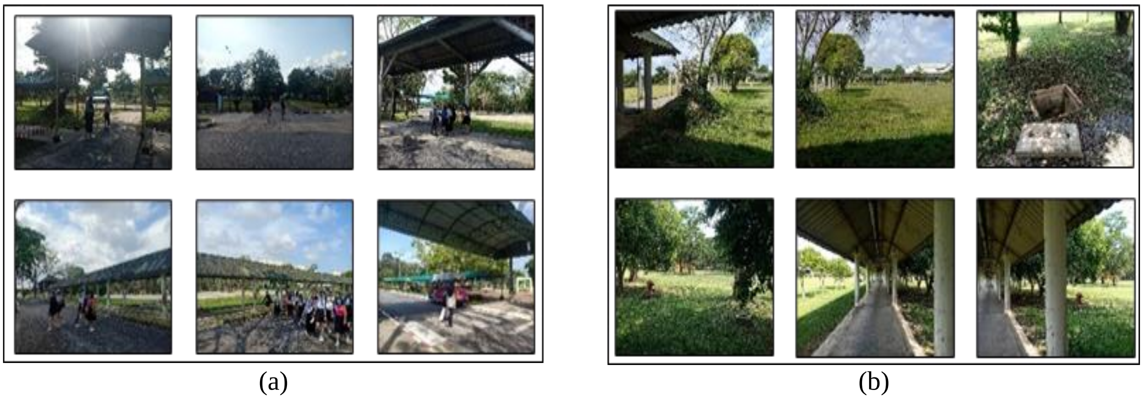

(a) (b)

Figure 6. Illustration of the CVC05 partially occluded pedestrian dataset: (a) examples of the positive (pedestrian) samples and (b) examples of the negative (non-pedestrian) samples

The WU dataset, proposed in this paper, was created with three phases: morning, afternoon, and evening respectively. Pedestrian images were initially collected with the following properties: a variation of appearance, a variety of postures, action, illumination, and diversity of occlusion labels. During the data collection, there were different groups of students: three persons, ten persons, twenty persons, and crowded persons crossed through the roads inside the campus area. Images, especially pedestrians were taken on both mobile phone resolution (4312×5760 pixels) and camera resolution (1280×720 pixels) in the dormitory area, academic building area, and the urban scene of Walailak University. Moreover, non-pedestrian images were collected around the campus road areas and campus buildings. Students from Innovation of Medical Informatics (2$^{nd}$ year) and the other department volunteered in this dataset, and all of the volunteer students wore masks in accordance with the COVID-19 pandemic's regulations. The dataset contains 9,691 training images and 2,727 testing images, comprising a total of 3,606 positive images and 8,812 negative images. The dataset generated during this study are not publicly available but are available from the corresponding author on reasonable request and with permission of Walailak University. Some sample data of the WU dataset is illustrated in Figure 7. Examples of the positive (pedestrian) samples can be seen in Figures 7(a) and 7(b) explain the examples of the negative (non-pedestrian) samples.



(a) (b)

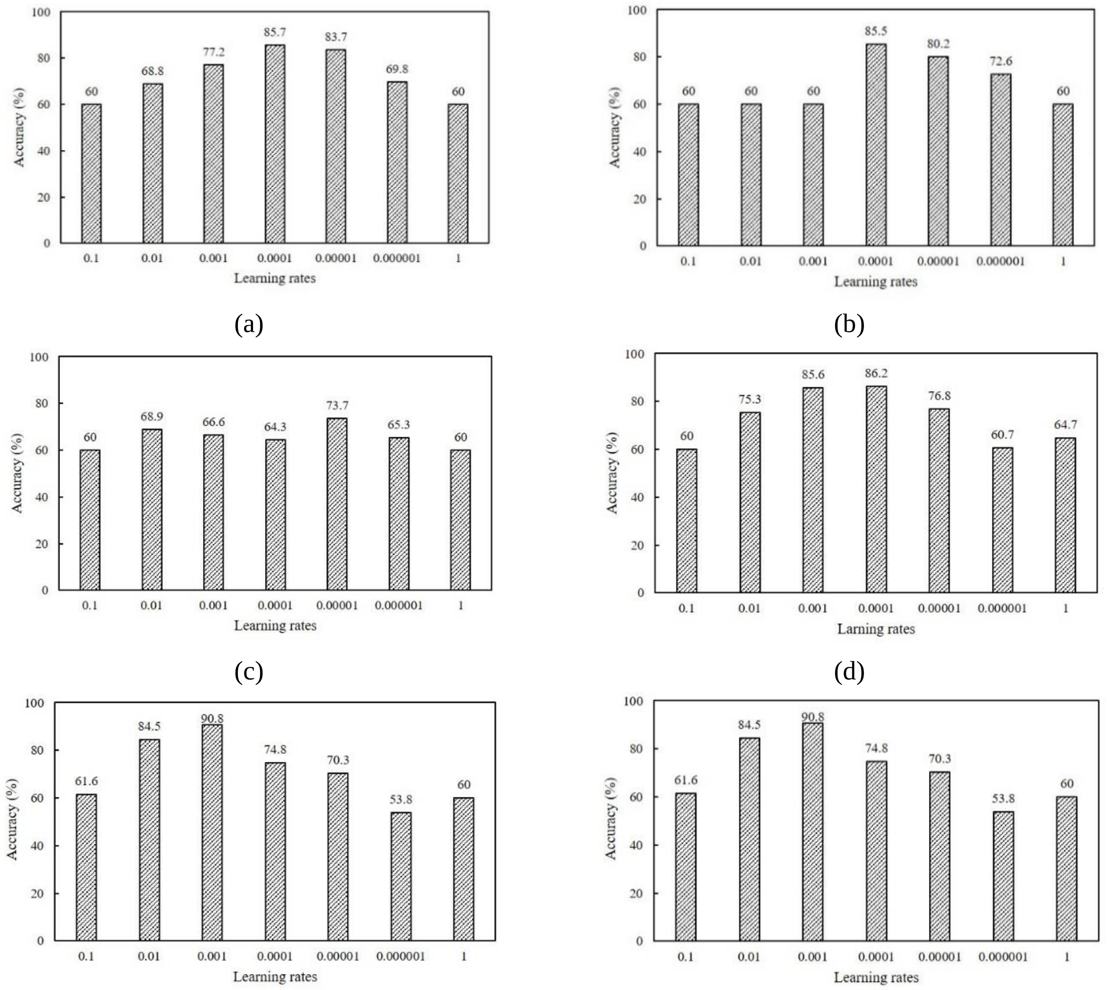
Figure 7. Illustration of the WU pedestrian dataset: (a) examples of the positive (pedestrian) samples and (b) examples of the negative (non-pedestrian) samples

### 4.2. Performance evaluation on end-to-end CNN model

In this experiment, we conducted to analyze the accuracy of the end-to-end CNN models with different types of optimizers: Adam, stochastic gradient descent (SGD), root mean square propagation (RMSprop), Adamax, Adagrad, and Adadelta by using the different number of learning rates. The optimizers are used to tune the parameters of a neural network in order to minimize the cost function. Furthermore, the appropriate selection of the optimizer is a critical factor in distinguishing between good training and inadequate training. In terms of regulating the weight loss and computing the gradient, each optimizer has advantages and disadvantages that vary according to learning rates and datasets. The purpose of analyzing the various learning rates is to determine how fast or slow the optimal weights should be calculated while calculating the gradient for the entire dataset. This experiment aims to find out the suitable optimizer and learning rate for solving the specific problems of the trained dataset. Figure 8 shows the comparison performance results for CNN hyper-parameter tuning.

The range of learning rates affects the performance accuracy of the CNN model using six types of optimizers. The studies were conducted out using the INRIA person dataset. In the training and testing samples, the entire datasets were mixed with diverse properties: view, appearance, and occlusion of pedestrians. The performance accuracy of the CNN model with Adam optimizer performed better on the learning rate with 0.0001, as shown in Figure 8(a). When the learning rates were low and the initial time steps were short, the classification scores declined by roughly 20%. The classification accuracy adopting Adam optimizers is 74.2%, according to the overall performance statistics.

The analysis of the performance results on the SGD optimizer is shown in Figure 8(b). The experiment results indicate that the CNN model achieved an accuracy of 85.5% when trained with the SGD optimizer at a learning rate of 0.0001. The initial learning rate climbed consistently during the optimal learning rates and then decreased by nearly 25% until the significant steps. Furthermore, the CNN model's average performance accuracy using SGD is 68.33%, which is 10% lower than the Adam optimizer. The analysis of the classification results on RMSprop is shown in Figure 8(c). The overall results of the CNN model attained 65.54%, and the performance accuracy reached 73.7% while using the optimal learning rates of 0.00001. Figures 8(d), 8(e), and 8(f) show the performance analysis of adaptive gradient optimization algorithms Adamax, Adagrad, and Adadelta, respectively. While training with learning rates of 0.0001 on Adamax, 0.001 on Adagrad, and 0.1 on Adadelta, the overall classification accuracy of three optimizers was around 70%. When compared to all of the optimizers, Adadetla achieved better classification results when employing small and large learning rates (0.1 and 1). According to the results of the experiment, a low learning rate demands several updates before reaching the minimum point, resulting in the lowest performance accuracy. The larger learning rate, on the other hand, skips the optimal point and increases the loss function. By comparing the result of all the experiments, it is concluded that the optimal learning rate of 0.00001 is the optimal choice for Adam optimizer training since it improves performance classification on the INRIA dataset and decreases the loss function.



(a) (b)

(c) (d)

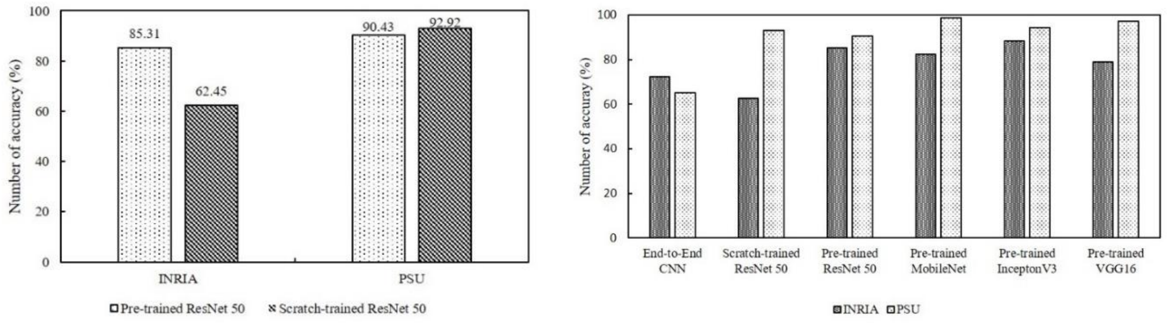
(e) (f)

Figure 8. Analysis of accuracy results for CNN Hyper-parameter tuning: (a) performance accuracy of CNN model with Adam; (b) performance accuracy of CNN model with SGD; (c) performance accuracy of CNN model with RMSprop; (d) performance accuracy of CNN model with Adamax; (e) performance accuracy of CNN model with Adagrad; and (f) performance accuracy of CNN model with Adadelta on the different number of learning rates

### 4.3. Deep feature extraction model on INRIA and PSU dataset

To overcome the partial occlusion handling problem in deep feature extraction models, the ResNet50 architecture is entirely scratch-trained on INRIA and PSU datasets. The consistency of the pre-trained ResNet50 and the scratch-trained ResNet50 for classifying the partially occluded regions of pedestrians was examined in this experiment. We used 224×224 input image size, the optimal hyper-parameter using Adam optimizer, and a learning rate of 0.00001 for training deep learning models, as mentioned in the previous section. Figure 9(a) shows the performance comparison between the pre-trained ResNet50 and the scratch–trained ResNet50. Results of the experiments were performed with different behaviors of data samples in which single pedestrians, crowded pedestrians, and partially occluded parts of pedestrians. This experiment used two methods: pre-trained and scratch-trained on ResNet 50 architecture to analyze the classification performance for INRIA and PSU datasets. In PSU dataset, the experiment results show that the scratch- trained ResNet 50 reached higher classification performance than the pre-trained ResNet 50. While training with the PSU dataset contained crowded and partially occluded parts of pedestrians, the interesting point is that the classification performance of the PSU dataset is improved 30%, which is higher than the INRIA dataset performance. However, the pre-trained model trained with the INRIA dataset attained the lowest classification accuracy (62.45%). The experiments demonstrated that even if pre-trained models are trained with millions of images from the ImageNet dataset, fine-tuning provides less accurate classification results than the end-to-end feature extraction models, especially datasets with partially occluded pedestrians. It can be observed that ResNet 50 reached out better feature extraction and classification even though the training dataset contained partially occluded parts of pedestrians.

Figure 9(b) shows the comparison performance of different classification models of the research study on INRIA and PSU datasets. The different properties of the two datasets have different kinds of classification performance. The PSU dataset especially outperformed the classification performance to handle the partially occluded parts of pedestrians. The experiments on end-to-end CNN found that the classifier misclassified on pedestrians with attached objects and pedestrians with black colors shirts as non-pedestrians and difficult to classify on the occluded parts of pedestrians. The classifier incorrectly detected pedestrians and performed worse on PSU data samples than on INRIA data samples because the occluded parts were assumed to be noises. In Southern Thailand datasets, the scratch-trained ResNet50 model correctly recognized significant occluded parts of pedestrians, including crowded pedestrians in complicated scenarios. The rich feature extraction of the deep convolutional neural network offered additional feature information of the occluded pedestrians, as shown in Figure 9. It is observed that the more profound the network layers are; the more specific feature extraction patterns provide. We conclude from this section by observing that the feature extracted from the pre-trained ImageNet task provides the small specific classification patterns on the deeper layer of neural networks.



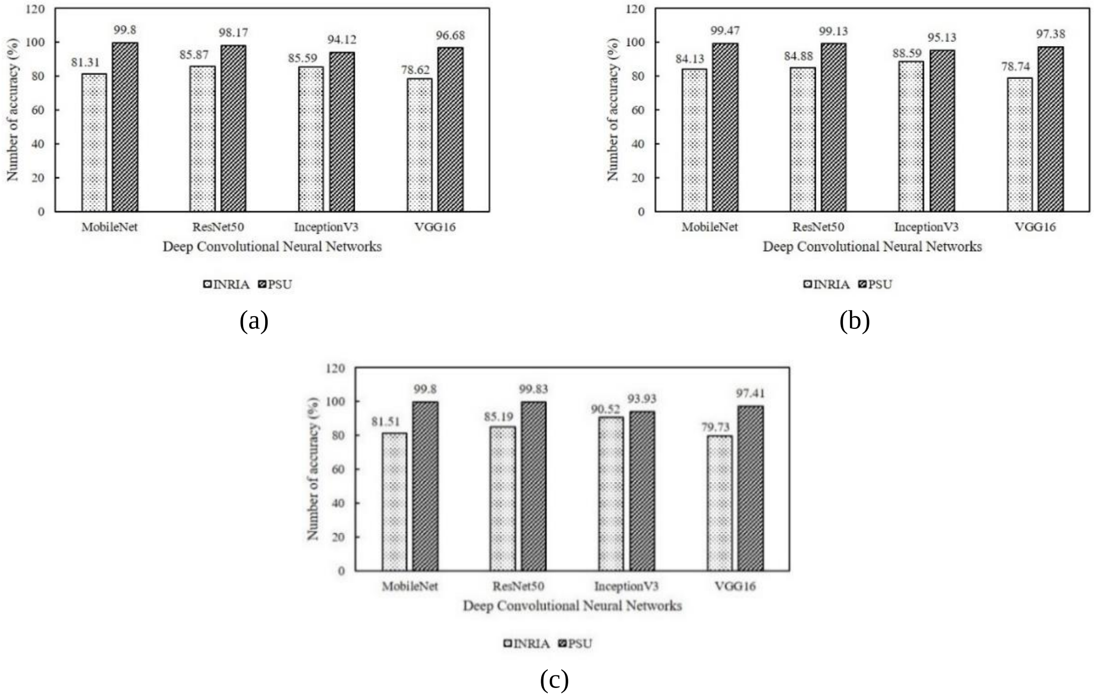
(a) (b)

Figure 9. Performance comparison of different classification models: (a) pre-trained ResNet50 and scratch-trained ResNet50 models and (b) different deep convolutional neural networks models

### 4.4. Pre-trained deep convolutional neural network model on INRIA and PSU dataset

Many hyper-parameters must be modified to train the CNN to be able to categorize images accurately; these hyper-parameters affect the network's performance during its time to convergence. This study used the different batch sizes with the optimal learning rates to train the pre-trained deep convolutional neural networks using INRIA and PSU datasets. The reason for using the different batch sizes is to investigate the impact on the learning process of CNN that converges quickly or slowly at the cost of noise in the training process. The pre-trained deep convolutional neural networks were MobileNet, ResNet50, InceptionV3, and VGG16. Figure 10 illustrate the performance comparison of the pre-trained models. Figure 10(a) shows the classification performance of four pre-trained models using the Adam optimizers of learning rate 0.00001 with batch size 8. In this experiment, the two datasets: INRIA and PSU were used to analyze the presence of the occluded parts in the pedestrian classification. As you can see in the results, the average classification performance of the PSU dataset on four models attained approximately 20% higher than the INRIA dataset. Compared with the four pre-trained models, MobileNet classification performance peaked at an accuracy (99.8%). It is observed that the advantages of the hierarchical feature representation of the deep convolutional networks helped to know about the patterns of the partially occluded parts of pedestrians while training with the presence of the occlusion data samples.

The pedestrian classification performance of the pre-trained models that are trained on INRIA and PSU datasets is described in Figure 10(b). The experiments were used with hyper-parameters as Adam optimizers of learning rates 0.00001 with batch size 16. In the INRIA dataset, the InceptionV3 pre-trained model increased with around 5% accuracy compared with the classification performance on batch size 8. Moreover, the overall average classification performance of four pre-trained models on the PSU dataset is 98%, which is approximately 14% higher than the INRIA dataset's performance. Even though the more partially occluded parts were used in training and testing data samples, the pre-trained filters were adapted to learn the valuable and comprehensible patterns of the data. In Figure 10(c), the performance comparison between different pre-trained networks is described and analyzed on the two datasets: INRIA and PSU. Four popular pre-trained models were used in this experiment to be accurate in the classification task on the ImageNet dataset. The classification performance of the PSU dataset reached nearly the entire percentage on classifying the partially occluded pedestrians compared with the INRIA dataset. In the PSU dataset, the average performance of four pre-trained models is around 97% accuracy, training with both occluded data samples and non-occluded data samples. Furthermore, the classification performance also achieved 13% higher than the INRIA dataset's performance (84%). As far as we study the deep neural networks, suitable hyper-parameter tuning is also needed to solve the specific problems depending on the amount of the data samples. Moreover, the MobileNet classification task is more stable on every batch size while training the optimal hyper-parameters.



(a)

(b)

(c)

Figure 10. Performance comparison between different pre-trained deep convolutional neural networks with Adam optimizers of learning rate (a) lr=0.00001 with batch size = 8, (b) lr=0.00001 with batch size = 16, and (c) lr=0.00001 with batch size = 32

### 4.5. Performance evaluation of different deep CNN models on CVC05 and WU pedestrian datasets

In this experiment, the performance comparison of the re-trained models has evaluated on both CVC05 partial occluded pedestrian dataset and WU pedestrian datasets. The reason why we choose CVC05 dataset to evaluate the different models is that CVC05 partial occluded pedestrian dataset mostly contained various illumination changes and thus some samples of the pedestrian were difficult to classify the foreground objects from the background. Additionally, the WU pedestrian dataset included information on different pedestrian postures, particularly those who appeared to be congested. Due to the fact that PSU training data had more partially occluded pedestrians than INRIA training data, two validation sets of pedestrian data were employed to evaluate on the various trained data samples. This is true for all the partially occluded pedestrians are challenging to classify in Table 3, thereby depending on the depth of the deep convolution neural network architectures. Comparing the architectures of deep convolutional neural networks yields varying performance accuracy and needs more training time, particularly for the deeper networks. As in the Table 3, we compare the quantitative outcomes of our six proposed models in precision (%mAP), with the two challenging partially occluded pedestrian datasets. In the end-to-end deep convolutional neural network model, the fifteen-layer CNN architecture achieved less performance (%mAP) on both CVC05 and WU pedestrian datasets. While utilizing the end-to-end deep convolutional neural network model on INRIA training data, the extracted features from the occluded regions are incapable of classifying pedestrians and are assumed to be noises (non-pedestrians). Otherwise, the pedestrian images with the attached objects would be impossible to manipulate in real-world contexts. However, the evaluation results for both datasets utilizing deep feature extraction models (scratch-trained ResNet50), pre-trained deep CNN model (ResNet50), and pre-trained deep CNN model (VGG16) increased by approximately 10% (mAP). An intriguing observation regarding the evaluation results is that the classifier performed well with partially occluded pedestrians when we included them in the training data. As evidenced by the performance of the pre-trained deep CNN model (InceptionV3), the evaluation accuracy of WU Val on PSU training data increased by about 20% (mAP) compared to INRIA training data. Meanwhile, improved evaluation results 77% (mAP) were also obtained on the partially occluded pedestrian dataset (CVC05), despite categorizing pedestrian images with dim illumination scenarios. More details regarding particular aspects of the pedestrian can be found on the more expansive the network architecture. Our proposed WU dataset produced remarkable assessment results 96% (mAP) using the PSU data samples in the pre-trained deep CNN model (MobileNet). It can be inferred that the addition of occluded pedestrian training instances provides the classifier with sufficient feature information to function well on complicated data samples in the real-world environment. In addition, the rich feature representation can enhance the classifier's ability to handle partially obscured portions under illumination. As we concluded from our analysis of the various deep CNN architectures, deeper neural network architectures perform better but take more time to train.

Table 3. Evaluation comparison results of two validation sets: CVC05 Val and WU Val on the end-to-end convolutional neural network model, deep feature extraction model, pre-trained VGG-16 model, pre-trained InceptionV3 model, pre-trained ResNet50 model, and pre-trained MobileNet model respectively. Training data: INRIA and PSU (Train-Test)

| Methods | Training Data | Val Samples | CVC05 Val %mAP | WU Val %mAP |
|---|---|---|---|---|
| End-to-End Convolutional Neural Network Model | INRIA | 2138 | 44 | 43 |
| Deep Feature Extraction Model (Scratch-trained ResNet-50) | INRIA | 2138 | 67 | 53 |
| Deep Feature Extraction Model (Scratch-trained ResNet-50) | PSU | 2138 | 66 | 53 |
| Pre-trained Deep CNN Model (VGG16) | INRIA | 2138 | 66 | 54 |
| Pre-trained Deep CNN Model (VGG16) | PSU | 2138 | 67 | 53 |
| Pre-trained Deep CNN Model (InceptionV3) | INRIA | 2138 | 77 | 55 |
| Pre-trained Deep CNN Model (InceptionV3) | PSU | 2138 | 78 | 72 |
| Pre-trained Deep CNN Model (ResNet50) | INRIA | 2138 | 67 | 53 |
| Pre-trained Deep CNN Model (ResNet50) | PSU | 2138 | 67 | 53 |
| Pre-trained Deep CNN Model (MobileNet) | INRIA | 2138 | 69 | 53 |
| Pre-trained Deep CNN Model (MobileNet) | PSU | 2138 | 77 | 96 |

## 5. CONCLUSION

The purpose of this study is to examine deep convolutional neural network analysis for pedestrian classification. The six classification models are utilized to implement and analyze pedestrian classification with variable hyper-parameter tuning to resolve partial occlusion handling. Additionally, the prediction accuracy of several convolutional neural networks is evaluated using pedestrian datasets from INRIA and PSU. This study aims to find out the classification performance of different networks on two datasets and analyze the consistency of the prediction by each network. In end-to-end convolutional neural network training, the different optimizers with different learning rates were used to find the suitable hyper-parameters for solving the occlusion problems of the trained dataset. It is observed that the small learning rate requires many updates before reaching the minimum point, resulting in the lowest performance accuracy. The pre-trained and scratch-trained ResNet50 architecture was utilized to examine different occlusion behaviors in both the INRIA and PSU datasets for deep convolutional neural network feature extraction. Even when pre-trained models are used to train using the massive amount of data samples from the ImageNet dataset, classification accuracy on the INRIA dataset is around 14% and 30% on the PSU dataset, which is lower than scratch-trained models' performance. However, the accuracy of the pre-trained MobileNet model is around at the peak of the accuracy (99.8%) on PSU dataset and (81.31%) on INRIA dataset.

Based on the aforementioned experiments, we can conclude that leveraging the knowledge of pre-trained techniques can enhance the adequate representation of feature information and may aid in understanding the complicated patterns of pedestrians. According to the research, a promising evaluation performance on the WU dataset was obtained by conducting a deeper convolutional neural network design (96%). Additionally, different neural network architectures have unique capacities for handling tasks and their particular information. In particular, crowded occluded pedestrian handling requires modification in order to provide an effective understanding of the feature information. Moreover, the hierarchical feature representation of the deep convolutional networks helped to know about the patterns of the partially occluded parts of pedestrians. Evaluation comparison results are shown in Table 3, where the deeper network architecture (MobileNet) outperformed the smaller network architecture (end-to-end convolutional neural network) in terms of classification accuracy (96%mAP) using the partially occluded pedestrian dataset. It is summarized from the experiments that the deeper network architectures provide more information of specific features and know the complex patterns of partially occluded parts of pedestrians. In the future, the combination of different features from the deep convolutional neural network will be used to develop the partial occlusion handling in pedestrian classification.

## ACKNOWLEDGEMENTS

This research was funded by the Institute of Research and Development, Walailak University, Grant No. WU65206. The author would also like to give the special thanks to Mr. Kan Rithy for encouraging and providing the research advice.

## REFERENCES

[1] K. Goniewicz, M. Goniewicz, W. Pawłowski, and P. Fiedor, "Road accident rates: strategies and programmes for improving road traffic safety," *European Journal of Trauma and Emergency Surgery*, vol. 42, no. 4, pp. 433–438, Aug. 2016, doi: 10.1007/s00068-015-0544-6.

[2] Y. Tanaboriboon and T. Sattiennam, "Traffic accident in Thailand," *IATSS Research*, vol. 29, no. 1, pp. 88–100, 2005, doi: 10.1016/S0386-1112(14)60122-9.

[3] J. Sukruay, "Road accident biggest health crisis," *Thailand Development Research Institute (TDRI Insight)*, 2020. https://tdri.or.th/en/2020/11/road-accidents-biggest-health-crisis/ (accessed Dec. 07, 2022).

[4] M. Thu and N. Suvonvorn, "Pyramidal part-based model for partial occlusion handling in pedestrian classification," *Advances in Multimedia*, vol. 2020, pp. 1–15, Feb. 2020, doi: 10.1155/2020/6153580.

[5] L. Ding, Y. Wang, R. Laganière, D. Huang, and S. Fu, "Convolutional neural networks for multispectral pedestrian detection," *Signal Processing: Image Communication*, vol. 82, p. 115764, Mar. 2020, doi: 10.1016/j.image.2019.115764.

[6] J. Gu, C. Lan, W. Chen, and H. Han, "Joint pedestrian and body part detection via semantic relationship learning," *Applied Sciences*, vol. 9, no. 4, p. 752, Feb. 2019, doi: 10.3390/app9040752.

[7] M. E. Mihçioğlu and A. Z. Alkar, "Improving pedestrian safety using combined HOG and haar partial detection in mobile systems," *Traffic Injury Prevention*, vol. 20, no. 6, pp. 619–623, Aug. 2019, doi: 10.1080/15389588.2019.1624731.

[8] E. J. Cheng *et al.*, "A fast fused part-based model with new deep feature for pedestrian detection and security monitoring," *Measurement*, vol. 151, p. 107081, Feb. 2020, doi: 10.1016/j.measurement.2019.107081.

[9] Y. Chen, H. Xie, and H. Shin, "Multi-layer fusion techniques using a CNN for multispectral pedestrian detection," *IET Computer Vision*, vol. 12, no. 8, pp. 1179–1187, Dec. 2018, doi: 10.1049/iet-cvi.2018.5315.

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, Apr. 2012, doi: 10.1109/TPAMI.2011.155.

[11] G.-S. Hong, B.-G. Kim, Y.-S. Hwang, and K.-K. Kwon, "Fast multi-feature pedestrian detection algorithm based on histogram of oriented gradient using discrete wavelet transform," *Multimedia Tools and Applications*, vol. 75, no. 23, pp. 15229–15245, Dec. 2016, doi: 10.1007/s11042-015-2455-2.

[12] T. Yamasaki, "Histogram of oriented gradients (HOG)," *The Journal of The Institute of Image Information and Television Engineers*, vol. 64, no. 3, pp. 322–329, 2010, doi: 10.3169/itej.64.322.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2004, vol. 1, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[14] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 32–39, doi: 10.1109/ICCV.2009.5459207.

[15] J. Zhu, S. Liao, Z. Lei, and S. Z. Li, "Multi-label convolutional neural network based pedestrian attribute classification," *Image and Vision Computing*, vol. 58, pp. 224–229, Feb. 2017, doi: 10.1016/j.imavis.2016.07.004.

[16] M. Sun, C. Wang, S. Wang, Z. Zhao, and X. Li, "A new semisupervised-entropy framework of hyperspectral image classification based on random forest," *Advances in Multimedia*, vol. 2018, pp. 1–27, Sep. 2018, doi: 10.1155/2018/3521720.

[17] W. Li, H. Ni, Y. Wang, B. Fu, P. Liu, and S. Wang, "Detection of partially occluded pedestrians by an enhanced cascade detector," *IET Intelligent Transport Systems*, vol. 8, no. 7, pp. 621–630, Nov. 2014, doi: 10.1049/iet-its.2012.0173.

[18] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: a review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019, doi: 10.1109/TNNLS.2018.2876865.

[19] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arxiv.org/abs/1409.1556*, Sep. 2014.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[22] A. G. Howard *et al.*, “MobileNets: efficient convolutional neural networks for mobile vision applications,” *arxiv.org/abs/1704.04861*, Apr. 2017.

[23] W. Ouyang *et al.*, “DeepID-Net: Object Detection with deformable part based convolutional neural networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1320–1334, Jul. 2017, doi: 10.1109/TPAMI.2016.2587642.

[24] Y. Tian, P. Luo, X. Wang, and X. Tang, “Deep learning strong parts for pedestrian detection,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1904–1912, doi: 10.1109/ICCV.2015.221.

[25] C.-Y. Lin, H.-X. Xie, and H. Zheng, “PedJointNet: joint head-shoulder and full body deep network for pedestrian detection,” *IEEE Access*, vol. 7, pp. 47687–47697, 2019, doi: 10.1109/ACCESS.2019.2910201.

[26] J. Xie, Y. Pang, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, “Mask-Guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3872–3884, 2021, doi: 10.1109/TIP.2020.3040854.

[27] D. Ribeiro, J. C. Nascimento, A. Bernardino, and G. Carneiro, “Improving the performance of pedestrian detectors using convolutional learning,” *Pattern Recognition*, vol. 61, pp. 641–649, Jan. 2017, doi: 10.1016/j.patcog.2016.05.027.

[28] P. Luo, Y. Tian, X. Wang, and X. Tang, “Switchable deep network for pedestrian detection,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 899–906, doi: 10.1109/CVPR.2014.120.

[29] S. Wang, J. Cheng, H. Liu, and M. Tang, “PCN: part and context information for pedestrian detection with CNNs,” *arxiv.org/abs/1804.04483*, Apr. 2018.

[30] X. Du, M. El-Khamy, J. Lee, and L. Davis, “Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection,” in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2017, pp. 953–961, doi: 10.1109/WACV.2017.111.

[31] K. Park, S. Kim, and K. Sohn, “Unified multi-spectral pedestrian detection based on probabilistic fusion networks,” *Pattern Recognition*, vol. 80, pp. 143–155, Aug. 2018, doi: 10.1016/j.patcog.2018.03.007.

[32] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in *ECCV 2016*, 2016, pp. 354–370.

## BIOGRAPHIES OF AUTHORS

**May Thu** received a B.E degree in Information Technology from Hmawbi Technology University, Hmawbi, Yangon, Myanmar, in 2012 and an M.E degree in Information Technology from Yangon Technology University, Yangon, Myanmar, in 2014. She is pursuing the Ph.D. degree in Computer Engineering from Prince of Songkla University, Hat Yai, Thailand. She is currently a Lecturer in the Department of Information Technology, School of Informatics, Walailak University (WU), Nakhon Si Thammarat, Thailand, and she is also a member of the Informatic Innovation Center of Excellence (IICE). Her current research interests include machine vision and image processing, complex data classification in pedestrian detection, human behavior analysis, supervised machine learning, and deep convolution neural network learning. She can be contacted at email: may.th@wu.ac.th.

**Nikom Suvonvorn** received Bachelor's degree of Computer Engineering (with Honorable), 1998 at Prince of Songkla University, Songkhla, Thailand, the Master's degree of Computer Engineering, in 2003 at ESME-Sudria, FRANCE, the Master's degree of Electronic System and Information Technology, DEA-SETI, Ecole Doctorale STITS, 2003 University of Paris XI, FRANCE, and Ph.D. (with Très Honorable), Institut d'Electronique Fondamentale, 2006 University of Paris XI, France. He is currently an Assistant Professor in the department of Computer Engineering at Prince of Songkla University and also a Director of the Intelligent Automation Research Center (IARC). His current research interests include computer vision and image processing for surveillance systems, human behavior analysis, elderly monitoring systems, software development, and distributed system. He can be contacted at email: kom@coe.psu.ac.th.

**Nichnan Kittiphattanabawon** received a bachelor's and Master's degree in Computer Science from Rangsit University in 1993 and Prince of Songkla University in 1999, respectively, and a doctoral degree in Technology (International Program) from Sirindhorn International Institute of Technology, Thammasat University in 2012. She is currently an Assistant Professor in the Department of Information Technology, School of Informatics, Walailak University. Her current research interests include text mining, data mining, document relation discovery, and association analytics. She can be contacted at: knichcha@wu.ac.th.