國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

使用 Swin Transformer 進行針對遮擋行人的行人偵測任務

Swin Transformer for Pedestrian and Occluded Pedestrian Detection

梁榮恩

Jung-An Liang

指導教授：丁建均 博士

Advisor: Jian-Jiun Ding, Ph.D.

中華民國 112 年 11 月

November 2023

doi:10.6342/NTU202304551

誌謝

當我在論文完成的這個重要時刻，我想要用這個機會，表達我最深切的感謝之情。在這段碩士生涯的過程中，有許多人給予我巨大的支持和鼓勵，讓我能夠克服種種挑戰，走到今天的這一步。

首先我要感謝我的父母。雖然親愛的父親已經離開了，但是先有父親的鼓勵與勸勉，我才敢鼓起勇氣報考了在職的碩士班。再來是母親在我求學的過程中，成了我的強力後援。在我忙碌的工作和課業中，打理並照顧我日常生活中的大小事，使我可以無後顧之憂地完成我的學業

再來要感謝我的摯愛的妻子，你一直是我的堅強後盾，支持我追求學術夢想。是你的理解、體諒、無條件的愛並且無怨無悔地照顧剛出世的女兒，使我能夠專心於研究和論文寫作，而無需擔心其他事情。你的存在讓我感到無比幸運，我永遠珍惜你。

對於我的指導教授，我也要深切的表達我的感激之情。老師的專業知識、耐心教導和各種啟發性的建議，在研究的過程中使我獲益良多。研究卡關的時候也是老師的耐心與包容，激勵我能夠不放棄，一直往前。

還要感謝召會中的弟兄姊妹，是你們一直以來的關心與問候，時刻在愛中的牧養照顧與成全，讓我能在心煩意亂，研究不順利的時候，能夠引導我回轉向神，尋求祂的幫助。

最後要感謝的就是在天上的神與祂的恩典。在整個研究和寫作的過程中，我充分地經歷了神的同在。祂所賜下的信心和能力一直伴隨著我，讓我能夠克服種種困難。這段過程中的每一步路都是神的恩典的展現，我深深感恩。

再次感謝所有支持和幫助過我的人，您們的愛和關懷是我寶貴的財富。這個論文的完成不僅是我的努力，也是您們的共同成就。我將永遠珍惜這段經歷，並將其視為我學習和成長的一個重要里程碑。

謝謝大家。

國立臺灣大學電機資訊學院電信工程學研究所 碩士生 梁榮恩

doi:10.6342/NTU202304551

中文摘要

本研究主要是提出一種高精確度的行人辨識模型。在自駕車所必須的道路物件辨識功能中，以辨識行人最為重要。因為與行人的碰撞事故一定傷亡最為嚴重，所以行人是車輛在道路上最不該與其發生碰撞的物件。本文預期後續能應用在自駕車的車內系統中，發揮即時偵測行人效果。因為於這幾年Transformer架構的模型在自然語言處理上的巨大成功，許多研究便將Transformer架構的模型試著應用在電腦視覺相關的任務上面。在其中的Vision Transformer中雖獲得了可以與CNN並駕齊驅的結果，但在訓練過程中其所需高額的運算量與龐大的模型參數量，對於要應用在終端的設備中，這是非常需要克服的難點。

2021年微軟所提出的Swin Transformer，其強大的性能、相比於Vision Transformer之下更為精簡的模型架構以及可在各種下游任務中廣泛且自由的應用，非常適合拿來當作一個物件偵測模型的特徵抽取器。本研究便利用其的強大功能來捕捉圖像中的多尺度特徵和空間關係，使其非常適合處理行人檢測這一具有挑戰性的任務。再搭配基於Faster R-CNN架構的兩階段的檢測器，結合階層式的RPN與ROI Head，並且在訓練RPN的過程中使用所有Anchors與Focal Loss。本研究在Euro City Persons和CityPersons數據集上的實驗展示了令人鼓舞的結果。特別是在檢測高度遮擋的行人方面，本研究的模型表現出色，展示了它應對傳統方法可能難以應對的挑戰性場景的能力。

關鍵字：深度學習、電腦視覺、物件偵測、Transformer

doi:10.6342/NTU202304551

ABSTRACT

This study primarily proposes a high-precision pedestrian recognition model. In the context of road object recognition, which is crucial for autonomous vehicles, pedestrian recognition holds paramount importance. Pedestrian collisions result in the most severe casualties, making pedestrians the objects that vehicles should avoid colliding with on the road. This paper is expected to be used in self-driving in-car systems to achieve real-time detection of pedestrians. Given the significant success of Transformer architecture models in natural language processing in recent years, researchers have explored the application of Transformer-based models in computer vision-related tasks. While the Vision Transformer (ViT) have shown promise in achieving results on par with Convolutional Neural Networks (CNNs), overcoming the high computational requirements and extensive model parameters during training poses a significant challenge, especially when targeting deployment on edge devices.

In 2021, Microsoft introduced the Swin Transformer, known for its powerful performance, more streamlined model architecture compared to the Vision Transformers, and its versatility in various downstream tasks. Swin Transformer is particularly suitable as a feature extractor for object detection models. This research harnesses its robust capabilities to capture multi-scale features and spatial relationships in images, making it well-suited for the challenging task of pedestrian detection. Combined with a two-stage detector based on the Faster R-CNN framework, which includes a cascade Region Proposal Network (RPN) and Region of Interest (ROI) Head, and the use of all anchors and Focal Loss during RPN training, this study showcases promising results in experiments conducted on the Euro City Persons and CityPersons datasets. Particularly,

doi:10.6342/NTU202304551

the model in this study demonstrates outstanding performance in detecting heavily occluded pedestrians, highlighting its ability to handle challenging scenarios that traditional methods may struggle with.

Keywords: Computer vision, Deep learning, Object detection, Transformer

doi:10.6342/NTU202304551

CONTENTS

doi:10.6342/NTU202304551

doi:10.6342/NTU202304551

LIST OF FIGURES

doi:10.6342/NTU202304551

LIST OF TABLES

doi:10.6342/NTU202304551

Chapter 1 Introduction

1.1 Research Background

In recent years, with rapid technological advancements, autonomous driving technology has become a prominent field in the automotive industry. The realization of autonomous driving holds significant importance in improving traffic safety, reducing accidents, and providing convenient means of transportation. However, in the process of achieving vehicle autonomy, road object recognition plays a crucial role. Among them, pedestrian recognition stands out as a key component since pedestrians are one of the most critical objects to avoid collisions with on the road. Therefore, the development of accurate and efficient pedestrian recognition algorithms is essential for realizing safe and reliable autonomous driving systems.

1.2 Research Objectives

The main objective of this study is to explore the application of pedestrian recognition techniques in vehicle autonomous driving systems. By analyzing existing methods, technologies, and algorithms for pedestrian recognition, this research aims to develop a precise and efficient pedestrian recognition system to enhance the safety and reliability of autonomous vehicles. Additionally, this study will investigate future trends in pedestrian recognition technology and propose corresponding improvements and optimizations.

1.3 Research Methodology

This study will employ a combination of experimental and analytical approaches. Firstly, a large dataset of pedestrian images and related data will be collected to establish a pedestrian database. Then, existing pedestrian recognition algorithms will be evaluated

doi:10.6342/NTU202304551

and compared to determine their strengths and limitations. Based on these analyses, novel pedestrian recognition algorithms will be developed, followed by detailed experiments, and testing to validate their performance and effectiveness. Finally, system optimizations and improvements will be conducted based on the experimental results, and relevant recommendations and conclusions will be presented.

1.4 Research Organization

This paper will consist of five main sections. The first section is the introduction, providing an overview of the research background, research objectives, research methodology, and the structure of the paper. The second section will review relevant n studies on pedestrian recognition, focusing on deep learning approaches. The third section will introduce the dataset and experimental design employed in this research. The fourth section will present the experimental results and discussions, analyzing the performance and effectiveness of the pedestrian recognition algorithms. Lastly, the fifth section will conclude and provide future perspectives, summarizing the research findings and proposing directions for future improvements and research.

In conclusion, this chapter has introduced the research background, research objectives, research methodology, and the structure of the paper. Pedestrian recognition is a crucial component in vehicle autonomous driving systems, significantly contributing to road safety. This study aims to develop an accurate and efficient pedestrian recognition system and explore the future trends in pedestrian recognition technology. The paper will validate the performance and effectiveness of pedestrian recognition algorithms through experiments and analysis, along with proposing relevant improvement methods.

doi:10.6342/NTU202304551

Chapter 2 Related Work

In the field of computer vision, object detection methods can now be classified into one-stage detectors and two-stage detectors. The main difference lies in whether predefined candidate boxes are used. Two-stage detectors typically include a region proposal network, which generates rough candidate boxes from the extracted feature map of an image. These candidate boxes are then fed into subsequent networks for object classification and more precise localization. On the other hand, one-stage detectors do not have this intermediate step and can directly perform object recognition tasks on the feature map. For example, in December 2015, Google introduced the Single Shot Detector (SSD) method, where the abstract of their paper begins with "We present a method for detecting objects in images using a single deep neural network."[1]

Two-stage detectors currently have a slight advantage over one-stage detectors in terms of object recognition accuracy, due to the presence of the region proposal network. However, because of the region proposal network, if multiple candidate boxes are selected, each of these selected boxes needs to be further processed by the subsequent neural network for classification and localization. As a result, two-stage detectors require more computational time and hardware resources compared to one-stage detectors. In the application of autonomous driving systems, the misidentification of pedestrians in pedestrian detection tasks can lead to severe consequences. Therefore, to achieve higher accuracy, this paper focuses on the methods of two-stage detectors.

doi:10.6342/NTU202304551

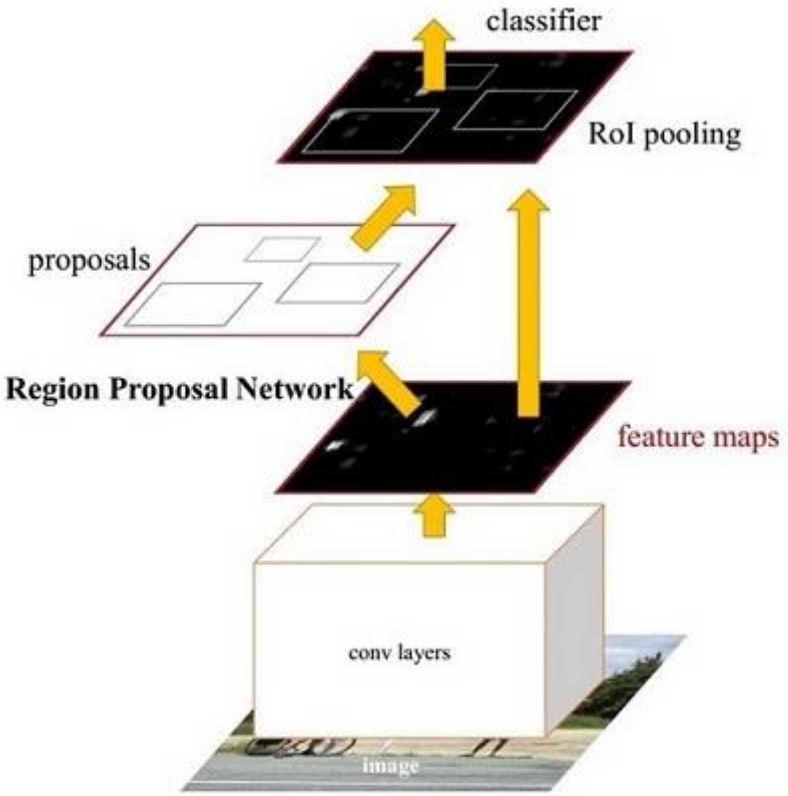


Fig 2.1 Architecture of Faster r-cnn [2]

Currently, the most well-known architecture among two-stage detectors is Faster R-CNN, which is based on the R-CNN framework[2]. The Faster R-CNN network architecture can be divided into three parts which as shown at Fig 1. The first part is the feature extractor, also known as the backbone, which extracts feature maps from the input image. The feature maps are then passed to the subsequent Region Proposal Network (RPN) network through Feature Pyramid Networks (FPN), which is referred to as the network's neck[3]. The second part is the Region Proposal Network (RPN). The goal of the RPN is to extract candidate boxes. The essence of object detection is to generate candidate boxes through regression. However, the network cannot generate candidate boxes of arbitrary sizes. Therefore, the RPN utilizes many anchors. These anchors, based on the feature map, divide the original image into many differently sized and aspect ratio rectangular boxes. The RPN performs rough classification and regression on these boxes,

doi:10.6342/NTU202304551

selecting a set of positive class boxes containing foreground objects and negative class boxes containing background regions after applying fine-tuning adjustments. These selected boxes are then fed into the subsequent network for training. The third part of Faster R-CNN is the ROI (Region of Interest) head. In the original Faster R-CNN, ROI pooling was used in this part, but it has been improved to ROI align in subsequent versions[4]. The ROI head takes the region proposals generated by the RPN and extracts fixed-sized feature maps for each region of interest. These feature maps are then fed into a fully connected network for object classification and bounding box regression. The ROI head is responsible for refining the localization and predicting the class labels of the proposed regions, ultimately producing the final detection results.

2.1 Faster RCNN Architecture

2.1.1 Feature Extractor

CNN Base. In the original Faster R-CNN paper, the backbone used is VGG16[5]. The VGG16 architecture is a deep convolutional neural network that consists of 16 layers, including convolutional layers, pooling layers, and fully connected layers. It is responsible for extracting high-level visual features from the input image, which are then used for region proposal and subsequent object classification and localization. However, recently VGG16 has been gradually replaced by ResNet in many applications. ResNet, short for Residual Network, is a deep convolutional neural network architecture that introduced the concept of residual connections[6]. These connections allow the network to learn residual mappings, making it easier to train very deep networks. Compared to VGG16, ResNet has shown better performance in terms of accuracy and gradient flow, especially in deeper network architectures. The popularity of ResNet in object detection tasks has grown due to its ability to effectively capture complex visual patterns and handle

doi:10.6342/NTU202304551

the challenges posed by deep networks. It has become a common choice as the backbone network in modern two-stage detectors, including Faster R-CNN variants. The integration of ResNet as the backbone has significantly improved the detection performance and contributed to the advancement of object detection research. In the field of pedestrian recognition, HRNet (High-Resolution Network) is a specialized type of convolutional neural network that has gained popularity due to its unique architecture and superior performance[7]. HRNet is specifically designed to handle the challenges posed by pedestrian images, such as variations in scale and the need to capture fine-grained details. Unlike traditional convolutional neural networks that downsample the input image, HRNet maintains high-resolution representations throughout its architecture. It achieves this by incorporating parallel branches that operate at different resolutions. This allows HRNet to simultaneously capture both high-level semantic information and fine-grained details, resulting in more accurate pedestrian recognition. HRNet adopts an information flow strategy that exchanges information across different resolution levels, enabling the network to effectively integrate multi-scale features. By preserving high-resolution representations, HRNet excels at capturing small-scale details that are crucial for accurate pedestrian recognition, such as facial features and clothing patterns. The architecture of HRNet consists of multiple stages, with each stage containing parallel sub-networks operating at different resolutions. The outputs from these sub-networks are fused together to form a comprehensive feature representation. This fusion of multi-scale features enables HRNet to capture both global context and local details, leading to improved accuracy in pedestrian recognition tasks. HRNet has achieved better performance on various pedestrian recognition benchmarks and competitions, surpassing previous approaches. Its ability to capture fine-grained details while maintaining a high level of semantic information makes it a valuable tool in the field. HRNet has found applications

doi:10.6342/NTU202304551

in real-world scenarios such as surveillance systems, autonomous vehicles, and human-computer interaction.

Transformer Base. Due to the outstanding achievements of the Transformers in the field of natural language processing, there has been an expectation that the Transformers could also be applied to image processing tasks. Several well-known Transformer models, such as Vision Transformer (ViT)[8] for image classification and Detection with Transformer (DETR)[9] for object detection, have shown promising results. However, they have not yet surpassed the dominance of convolutional neural networks (CNNs) in image processing. Since the introduction of the Swin Transformer[10] in 2021, it has disrupted the previous belief that Transformers cannot outperform CNNs in the field of computer vision. The Swin Transformer has emerged as a potential perfect alternative to CNNs. The lack of significant breakthroughs when applying Transformers to image processing tasks, as opposed to their success in natural language processing (NLP), can be attributed to two main reasons. Firstly, NLP and image processing involve different scales. NLP operates on fixed and standardized scales, while images exhibit a wide range of scale variations. This discrepancy in scale poses a challenge when directly applying Transformers to image data. Secondly, image processing tasks require higher resolution compared to NLP. Additionally, the computational complexity of Transformers in image processing is quadratic with respect to the image size, resulting in significant computational demands.

To address these challenges, Swin Transformer introduces two key improvements compared to previous models like Vision Transformer (ViT). Firstly, Swin Transformer adopts a hierarchical construction method similar to convolutional neural networks (CNNs). For example, it includes feature maps at different levels of down-sampling ratios, such as 4x, 8x, and 16x. This hierarchical backbone facilitates tasks like object detection

doi:10.6342/NTU202304551

and instance segmentation. In contrast, Vision Transformer directly down-samples by a factor of 16 from the beginning and maintains the same down-sampling rate for subsequent feature maps. Secondly, Swin Transformer incorporates the concept of Windows Multi-Head Self-Attention (W-MSA). For instance, at 4x and 8x down-sampling, the feature maps are divided into non-overlapping regions called windows, and Multi-Head Self-Attention is applied within each window. This approach reduces computational complexity, especially for large shallow feature maps, compared to Vision Transformer, which applies Multi-Head Self-Attention to the entire global feature map. However, this partitioning of feature maps into windows can limit information exchange between different windows. To address this issue, the Swin Transformer applied the concept of Shifted Windows Multi-Head Self-Attention (SW-MSA), which allows information to flow between adjacent windows. By incorporating these improvements, the Swin Transformer effectively tackles the challenges of scale variation and computational complexity, making it a promising choice for various computer vision tasks such as image classification, object detection, and semantic segmentation.

2.1.2 Feature Pyramid Networks

In most traditional object detection algorithms, only the top-level features are used for prediction. However, it is known that lower-level features have less semantic information but provide accurate object localization, while higher-level features contain richer semantic information but have coarser object localization. Although some algorithms incorporate multi-scale feature fusion, it is typically done on fused features for prediction. The key difference here is that in the concept of the feature pyramid[3], predictions are made independently at different feature levels. In other words, the feature pyramid addresses the trade-off between semantic information and positional accuracy by leveraging features at different scales. It recognizes that lower-level features excel at

doi:10.6342/NTU202304551

precise object localization, while higher-level features offer a deeper semantic understanding. By using a feature pyramid, the algorithm can make independent predictions at different feature levels. This enables capturing both fine details and high-level semantics simultaneously, resulting in improved object detection performance. The feature pyramid fully exploits the strengths of different feature levels, allowing the algorithm to better balance the relationship between object position and semantic information.

Subsequent to the introduction of the feature pyramid, several improvements have been proposed, such as PAFPN (Path Aggregation Network)[11], NAS-FPN (Neural Architecture Search-based Feature Pyramid Network)[12], and FPG (Feature Pyramid Grid)[13]. These methods aim to further enhance the effectiveness and efficiency of the feature pyramid for object detection. These advancements in feature pyramid methods have contributed to the ongoing progress in object detection, enabling more accurate and efficient detection of objects at various scales and aspect ratios.

2.1.3 Region Proposal Network

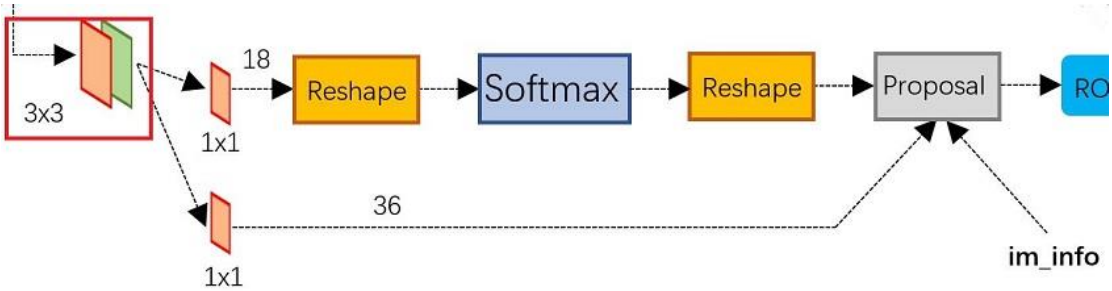


Fig 2.2 Architecture of RPN in Faster r-cnn [2]

Traditional methods are very time-consuming in generating detection boxes. For example, OpenCV adaboost uses sliding windows and image pyramids to generate detection boxes[14]. Or use the selective search method to generate detection boxes such as RCNN. Faster RCNN abandons the traditional sliding window and selective search

doi:10.6342/NTU202304551

methods, and directly uses RPN to generate detection boxes. This is a huge advantage of Faster R-CNN, which can greatly improve the generation speed of detection frames. Fig 2 shows the specific structure of the RPN network. It can be seen that the RPN network is actually divided into 2 lines. The upper one obtains positive and negative classifications through softmax classification anchors, and the lower one is used to calculate the bounding box regression offset for anchors to obtain an accurate proposal. The final proposal layer is responsible for synthesizing the positive anchors and the corresponding bounding box regression offset to obtain proposals, while eliminating proposals that are too small and beyond the boundary. When the entire network reaches the proposal layer, the target positioning is completed. Anchors play a fundamental role in RPN. Anchors are pre-defined bounding boxes of various sizes and aspect ratios that are overlaid onto an image. The purpose of anchors is to serve as reference boxes that anchor or localize potential objects in an image. These anchors act as starting points for generating region proposals during the object detection process. By covering a range of sizes and aspect ratios, anchors can capture objects of different scales and shapes. In the RPN, the anchors are generated based on a feature map obtained from the backbone network. The feature map is divided into a grid of cells, and each cell corresponds to a set of anchors of different sizes and aspect ratios. For each anchor, the RPN predicts the probability of it containing an object (foreground) or being background, as well as the adjustments needed to refine its position and size. During training, the RPN compares the predicted anchors with ground-truth bounding boxes to determine which anchors are positive (matching a ground-truth object) or negative (background). The positive anchors are used to compute classification and regression losses, which are then used to update the RPN's parameters during the training process. By using anchors, the RPN can generate a set of candidate region proposals that potentially contain objects of interest in an image. These proposals

doi:10.6342/NTU202304551

are subsequently refined and classified by the object detection network. Anchors help to localize and classify objects accurately by providing a structured set of bounding box priors that guide the object detection process.

In the existing RPN framework, there are still some challenges to overcome. For instance, to obtain sufficient positive samples, a set of anchors with different scales and ratios is required. Setting appropriate scales and aspect ratios is crucial for the performance of object detection, and it often requires extensive manual tuning. To address this issue, the Cascade RPN paper proposes a two-stage RPN with adaptive convolution to automatically adjust and align anchors with the targets[15]. In the Cascade RPN approach, the paper systematically tackles the problems of the RPN. Firstly, in the first stage, a single anchor is used instead of using multiple scales and aspect ratios. The definition of positive sample regions combines both anchor-based and anchor-free evaluation criteria to improve detection performance. This approach replaces the use of multiple scales and aspect ratios of anchors. Secondly, after each stage, adaptive convolution is applied to refine the anchors. This adaptive convolution maintains the alignment between anchor boxes and features while gaining refinement benefits from multiple stages. It ensures that the refined anchors remain properly aligned with the underlying features, leading to improved detection accuracy. By incorporating these techniques, the Cascade RPN method effectively addresses the challenges of the RPN. It simplifies the anchor design by using a single anchor and improves the detection performance by combining anchor-based and anchor-free evaluation criteria. Additionally, the adaptive convolution preserves the alignment between anchors and features throughout the refinement process, contributing to enhanced object detection accuracy.

2.1.4 ROI Heads

Faster R-CNN's ROI (Region of Interest) heads are an important component of the architecture. Once the region proposals are generated by the RPN (Region Proposal Network), the ROI heads refine and classify these proposed regions. The ROI heads consist of two main parts: the ROI pooling layer and the fully connected layers. The ROI pooling layer takes the region proposals and extracts fixed-sized feature maps from the corresponding feature map obtained from the backbone network. This allows for consistent feature map sizes for subsequent processing. The fixed-sized feature maps from the ROI pooling layer are then fed into a series of fully connected layers. These layers perform two tasks: classification and regression. In the classification task, the network assigns a probability score to each region proposal for different object categories, indicating the likelihood of containing a specific object class. The regression task adjusts the coordinates of the bounding box for each proposal, refining its position and size to better align with the object boundaries. By incorporating the ROI pooling layer and fully connected layers, the ROI heads enable accurate object classification and precise bounding box regression for each proposed region. This helps improve the overall object detection performance of the Faster R-CNN model. However, ROI pooling suffers from a limitation where it can only map the ROIs to the grid cells of the feature map at integer precision, resulting in information loss and positional misalignment.

To overcome the limitations of ROI pooling, ROI align was introduced as an improvement over ROI pooling[4]. ROI align uses bilinear interpolation within each sub-region of the ROI, allowing for mapping of the coordinates of the region to floating-point coordinates on the feature map, enabling more accurate position alignment and finer feature extraction. Specifically, ROI align divides the ROI into smaller sub-regions and performs bilinear interpolation within each sub-region to obtain feature values at floating-point coordinates. The feature values from these sub-regions are then merged to form the output feature representation of ROI align. Compared to ROI pooling, ROI align provides

doi:10.6342/NTU202304551

more accurate position alignment and finer feature representations. This improvement helps enhance the accuracy and localization precision of object detection models, particularly for small objects or cases that require more precise detection. As a result, many state-of-the-art object detection models have adopted ROI align as a replacement for ROI pooling.

In Cascade R-CNN[16], one of the improvements in the ROI heads addresses the issue of low-quality proposals generated by the Region Proposal Network (RPN). These low-quality proposals make it challenging to directly use a high threshold detector. To overcome this, Cascade R-CNN introduces cascade regression as a mechanism for re-sampling proposals and progressively increasing their IoU values across stages. Each stage in Cascade R-CNN utilizes a ROI head that is trained with a specific IoU threshold. This threshold is gradually increased across stages, allowing the ROI head in each stage to focus on a subset of proposals that meet the higher threshold. By doing so, the cascade regression helps improve the quality of proposals, making them better suited for subsequent stages with higher IoU thresholds. Cascade R-CNN addresses the issue of low-quality proposals by using cascade regression as a re-sampling mechanism. By progressively increasing the IoU values of proposals across stages, Cascade R-CNN improves the matching between proposals and detectors with higher IoU thresholds, leading to better object detection performance.

2.2 Loss Function

The choice of the loss function is crucial during the training process of the model. In the architecture of Faster R-CNN, different stages require the use of different loss functions, and the model's weights are updated by summing all the losses. Typically,

doi:10.6342/NTU202304551

classification loss, such as cross-entropy, is used to determine the class of the image. This loss measures the discrepancy between the predicted class probabilities and the true class labels. It guides the model to accurately classify objects in the image. For precise bounding box localization, regression loss, such as L1 loss (mean absolute error), is commonly employed. This loss measures the difference between the predicted bounding box coordinates and the ground truth coordinates. It helps the model to learn the precise location and size of the objects. During training, the losses from different stages, including the classification loss and regression loss, are summed to obtain the total loss. The model's weights are then updated using backpropagation and gradient descent based on this total loss. This process iterates until the model converges and achieves better performance in classifying objects and localizing bounding boxes.

Focal loss is a specialized loss function that was introduced to address the problem of class imbalance in object detection tasks[17]. In scenarios where the number of background (negative) samples significantly outweighs the number of foreground (positive) samples, traditional loss functions like cross-entropy may struggle to effectively train the model. The key idea behind focal loss is to assign higher weights to hard-to-classify examples, which are typically the minority class samples. It achieves this by introducing a modulating factor called the focal parameter. The focal parameter reduces the loss contribution of well-classified examples and increases the loss contribution of misclassified examples. Focal loss has shown significant improvements in object detection tasks, particularly in scenarios with imbalanced datasets. It effectively addresses the problem of dominant background samples and helps the model to better learn and distinguish the foreground objects of interest.

L1 and L2 loss in bounding box regression treat the four points of the bounding box independently and calculate the loss by summing the errors of each coordinate. They do

doi:10.6342/NTU202304551

not consider the correlation between these coordinates. On the other hand, IOU loss, which is based on the intersection over union metric, considers the correlation between bounding box coordinates and provides a more comprehensive evaluation of the predicted box compared to the ground truth box[18]. IOU loss measures the overlap between the predicted box and the ground truth box by calculating the intersection area and the union area. It captures the spatial relationship between the coordinates of the bounding boxes and penalizes predictions with lower IOU values. This allows the model to learn the spatial correlations and make more accurate predictions in terms of box localization and shape. Several improvements to IOU loss have been proposed to address its limitations. GIOU (Generalized IOU) loss extends IOU by incorporating the notion of the smallest enclosing box[19]. It not only measures the overlap between predicted and ground truth boxes but also considers the bounding box's size and spatial alignment. By penalizing both the localization error and the inconsistency in box sizes, GIOU loss provides a more comprehensive evaluation metric. DIOU (Distance IOU) loss introduces an additional term that measures the distance between the centers of predicted and ground truth boxes[20]. By including this distance term, DIOU loss encourages better alignment of the predicted box with the ground truth box in terms of both overlap and center proximity. CIOU loss (Complete IOU) extends DIOU loss by incorporating an aspect ratio term that measures the similarity in shape between the predicted and ground truth boxes[21]. This aspect ratio term helps to penalize predictions with significant distortions compared to the ground truth shape.

The high loss values associated with low-quality samples can cause excessively large gradients, leading to unstable parameter updates during model training. Focal IOU loss addresses this issue by penalizing the high loss caused by low-quality samples, effectively reducing the impact of these samples on the gradients[22]. By incorporating the focal

doi:10.6342/NTU202304551

mechanism into the IOU loss calculation, Focal IOU assigns lower weights to low-quality samples, reducing their contribution to the overall loss. This weight adjustment helps mitigate the issue of excessively large gradients caused by high loss values and allows for more stable parameter updates. As a result, the training process becomes less prone to drastic parameter changes, and the model can converge more effectively towards better performance. Focal IOU loss provides a means to alleviate the instability caused by high loss values from low-quality samples, allowing for smoother training dynamics and improved convergence during the training process.

2.3 Sampler

In object detection tasks, negative samples, typically referred to as background samples, are samples that do not contain any objects of interest. However, some negative samples may exhibit certain similarities with the objects, such as having similar textures or shapes. These challenging negative samples can have a negative impact on the model training and convergence, as the model may mistakenly identify them as positive samples. n traditional training processes, due to the large number of negative samples compared to positive samples (samples containing objects), the model may tend to classify them as negative and overlook some challenging negative samples. This can lead to a performance drop when the model encounters backgrounds that are similar to the objects of interest. Online Hard Example Mining (OHEM) is a technique designed to address this issue by selecting hard samples during training[23]. Hard samples refer to those samples with high loss values in classification or regression tasks, specifically the background samples that are similar to the objects. By selecting hard samples for training, the model is forced to learn how to better differentiate between these challenging samples and the objects, thus improving its accuracy and generalization ability when dealing with challenging negative

doi:10.6342/NTU202304551

samples. In the Faster R-CNN architecture, Online Hard Example Mining (OHEM) can be applied to improve the training process by selecting hard negative samples during each training iteration.

Region Proposal Network (RPN) Training. Train the RPN module to generate region proposals. During this process, both positive and negative samples are selected based on their IoU (Intersection over Union) overlap with the ground truth boxes. However, instead of randomly sampling negative samples, OHEM can be applied to select hard negative samples, which are challenging backgrounds that are similar to the objects of interest.

ROI Head Training. Train the ROI head module to perform object classification and bounding box regression based on the extracted features. During this stage, OHEM can be applied again to select hard negative samples to improve the model's ability to distinguish challenging backgrounds from actual objects.

2.4 Non-Maximum Suppression

Faster R-CNN utilizes Non-Maximum Suppression (NMS) to refine the output of the object detection process and remove redundant bounding box predictions. First, The RPN generates a set of region proposals by scanning the image at multiple scales and aspect ratios. Each proposal is associated with a confidence score indicating the likelihood of containing an object. The region proposals are passed through the ROI align layer, which extracts fixed-sized feature maps. These features are then fed into the ROI head, where object classification and bounding box regression are performed. For each region proposal, the model predicts the probability of object presence and refines the bounding box coordinates. The region proposals are filtered based on a confidence score threshold. Only the proposals with confidence scores above a certain threshold are considered for further processing, while the rest are discarded as low-confidence predictions. NMS is

doi:10.6342/NTU202304551

applied to the remaining region proposals to eliminate duplicate and overlapping predictions. The selected proposals after NMS represent the most confident and non-overlapping predictions. These proposals, along with their associated class labels and refined bounding box coordinates, are considered as the final output of the Faster R-CNN model. By applying NMS, Faster R-CNN effectively eliminates redundant detections and retains only the most confident and distinct bounding box predictions. This improves the accuracy and efficiency of object detection by reducing duplicate outputs and simplifying the downstream analysis.

2.5 Normalization

The limitation with Batch Normalization (BN) arises when the available computational resources or memory constraints restrict the model to use only a small batch size. In such cases, the effectiveness of BN may be compromised. When the model is limited to a small batch size, BN may not accurately estimate the statistics needed for normalization, leading to suboptimal performance. This is because BN computes the statistics (mean and variance) based on the statistics of the mini-batch, and with a small batch size, the estimated statistics may not be representative of the entire dataset. To address this issue, Group Normalization (GN) was proposed as an alternative to BN. GN divides the channels of a feature map into groups and computes normalization statistics within each group independently[24]. This is in contrast to BN, which computes statistics across the entire batch. By using group-wise normalization, GN reduces the dependency on the batch size and performs normalization effectively even with smaller batch sizes.

doi:10.6342/NTU202304551

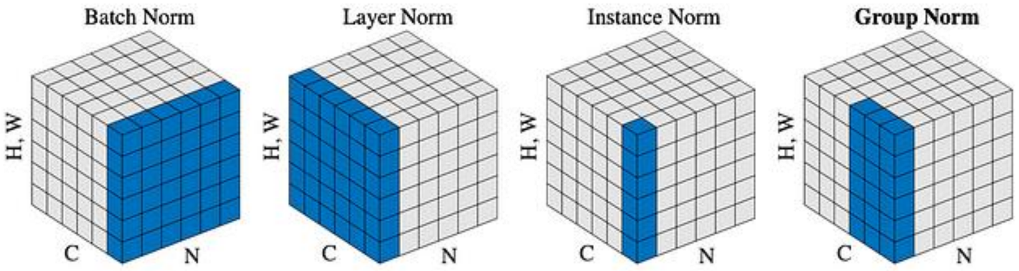


Fig 2.3 Comparison of different Norms [24]

By decoupling the normalization from the batch dimension and performing it within groups, Group Normalization reduces the reliance on large batch sizes and improves the model's performance, especially when dealing with limited computational resources or large input image sizes. It has been shown to be effective in various tasks and has gained popularity as an alternative to Batch Normalization in scenarios where small batch sizes are preferred or necessary.

Weight Standardization is an alternative technique to normalize the weights of neural network layers[25]. It aims to address the limitations of Batch Normalization and Group Normalization by normalizing the weights themselves rather than normalizing the activations. In traditional normalization techniques like BN or GN, the mean and standard deviation are computed using the activations within a mini-batch or groups of channels. However, Weight Standardization directly normalizes the weights of each layer by subtracting the mean and dividing by the standard deviation of the weight values. By normalizing the weights, Weight Standardization helps to stabilize the learning process and improve the generalization performance of the model. It has been observed that Weight Standardization can alleviate the issues caused by vanishing or exploding gradients and reduce the sensitivity to the initialization of network weights. It also enables the use of larger learning rates during training. Compared to BN and GN, Weight Standardization has the advantage of not relying on mini-batch statistics or groups of

doi:10.6342/NTU202304551

channels, making it applicable to various batch sizes and network architectures. It has been shown to be particularly effective in scenarios with small batch sizes or when the data distribution varies significantly across samples. Overall, Weight Standardization is a normalization technique that operates directly on the weights of neural network layers. It offers an alternative to traditional activation-based normalization methods like BN and GN, providing improved stability, generalization, and applicability across different network setups.

Chapter 3 Proposed Approach

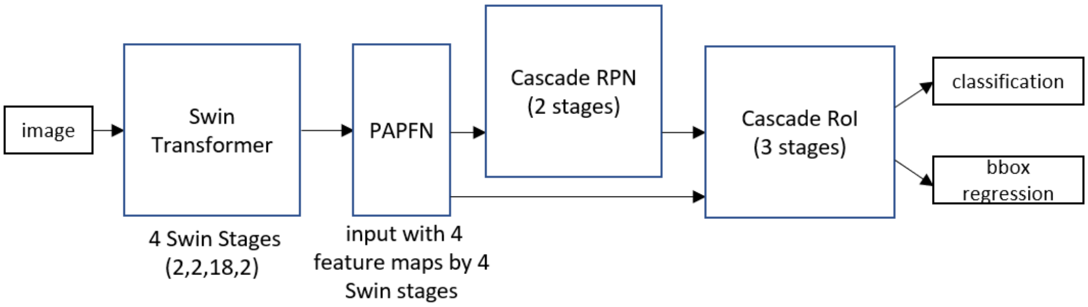


Fig 3.1 Architecture of our model

Pedestrian detection is a task in computer vision that aims to recognize and locate pedestrians in images or videos. In this task, other objects are considered as background, and the algorithm focuses solely on pedestrians. By employing image processing and machine learning techniques, pedestrian detection plays a crucial role in various applications such as traffic safety, video surveillance, and autonomous driving.

Because the two-stage detector can achieve more accurate localization, the entire model architecture is designed using a two-stage framework which is shown as Fig 3.1. It includes a feature extractor, an RPN (Region Proposal Network), and an ROI Align connected to the final ROI Head. We believe that with this architecture, we can achieve better performance in pedestrian detection. In this two-stage architecture, the feature extractor is first used to extract features from the input image. Then, the RPN is responsible for generating candidate pedestrian bounding boxes. The RPN uses anchors to generate candidate boxes at different positions and scales. It classifies and performs bounding box regression based on the features within each box, filtering out candidate boxes that potentially contain pedestrians. Next, the RPN-generated candidate boxes are aligned with the feature map using ROI Align, which extracts local features required for each candidate box. ROI Align addresses the issue of pixel misalignment in traditional

doi:10.6342/NTU202304551

ROI Pooling methods, thus improving localization accuracy. Finally, the ROI Head performs classification and bounding box regression on each candidate box. The ROI Head typically consists of fully connected layers and classifiers, which classify the extracted local features to determine whether a candidate box contains a pedestrian and perform further refinement of the bounding box.

3.1 Feature Extractor

In the feature extractor, we have chosen the Swin Transformer v2 as the backbone model, which has shown excellent performance in recent years. Pedestrians on the road exhibit diverse poses and may be occluded by objects or blend with the surrounding background, making pedestrian recognition challenging. Therefore, we utilize the Swin Transformer, which can balance computational complexity while incorporating the entire input image through window shifting self-attention mechanisms. The Swin Transformer has demonstrated outstanding results in various computer vision-related papers and downstream tasks, such as object classification, object segmentation, semantic segmentation, and more. Its ability to capture global contextual information and handle long-range dependencies makes it well-suited for addressing the challenges posed by pedestrians in complex scenes. Swin Transformer v2 is an improved version of the Swin Transformer, with three main enhancements[26]:

Post-normalization: The v2 version incorporates layer normalization after the self-attention layer and MLP (Multi-Layer Perceptron) block. This normalization helps stabilize gradients and improves the overall training stability.

Scaled cosine attention approach: Instead of using dot product similarity, the v2 version employs the cosine similarity measure to compute the relationships between token pairs. In the original self-attention mechanism, the similarity measurement between pairwise

doi:10.6342/NTU202304551

features is computed using the dot product. However, the authors of the Swin Transformer v2 observed that when replaced with post-normalization, in larger models, certain blocks or heads may be dominated by specific features in the attention map. To address this issue, they replaced the dot product similarity with cosine similarity. By using cosine similarity, they aim to mitigate the problem of certain feature pairs having excessively large dot products, which can dominate the attention computation. The cosine function inherently ranges between -1 and 1, and it is already normalized. This normalization property helps alleviate the issue of a few dominant features overpowering the attention mechanism.

Log-spaced continuous position bias: The v2 version redefines the relative position encoding by introducing log-spaced continuous position bias. This helps the model better capture positional information and enables more effective handling of long-range dependencies.

By incorporating these improvements, the Swin Transformer v2 aims to mitigate the limitations of its predecessor and enhance its performance, stability, transferability, and memory efficiency.

3.2 Feature Pyramid Networks

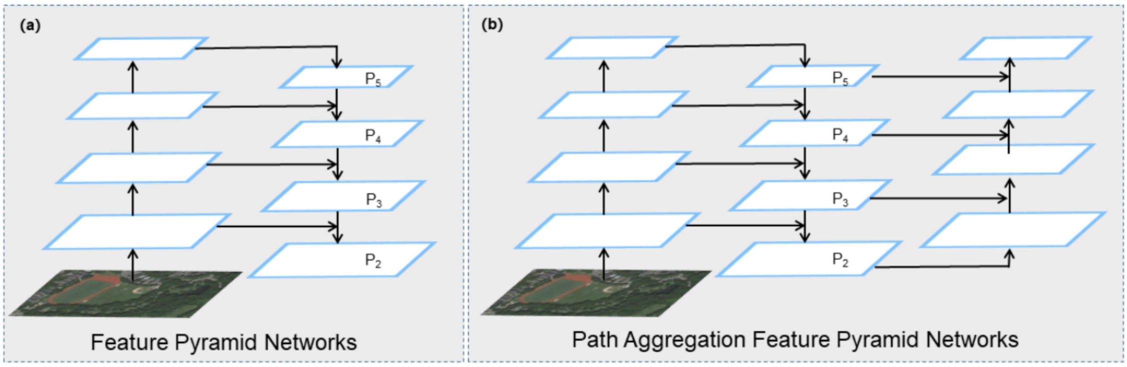


Fig 3.2 (a) Structure diagram of FPN and (b) Structure diagram of PAFPN. Comparing the two, PAFPN has one more bottom-up path than FPN[27]

doi:10.6342/NTU202304551

After conducting extensive experiments, we have found that PAFPN (Path Aggregation Network) yields the best performance improvement for pedestrian recognition. Therefore, we have chosen PAFPN as the FPN (Feature Pyramid Network) in our model. In contrast, other FPN architectures such as NASFPN or DYHead did not enhance the overall recognition performance and, in fact, resulted in a negative impact.

According to the original research papers, NASFPN[12] is specifically designed for RetinaNet[17], a single-stage object detection framework. NASFPN aims to optimize the feature pyramid network architecture for the RetinaNet model. Single-stage detectors lack the RPN and ROI align and directly predict bounding boxes after obtaining feature maps from the feature extractor. Due to the structural differences between single-stage and two-stage detectors, applying FPNs that perform well in single-stage detectors directly to two-stage detectors can have adverse effects.

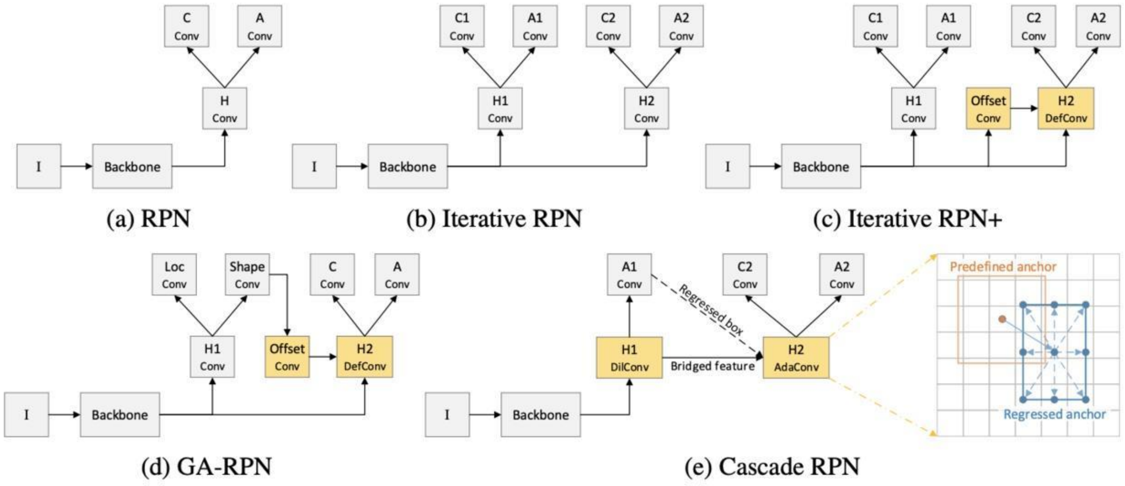
On the other hand, DyHead[28] is designed to be added after the FPN, utilizing neural networks to enhance semantic structure. However, since DyHead adds several blocks of neural networks, it incurs additional computational resources. The original paper suggests that stacking 6 blocks of DyHead yields the best results. Stacking 6 blocks of DyHead is not affordable or feasible for our research due to the increased computational cost and additional GPU memory consumption. It is not a matter of practicality, but rather a limitation in terms of available resources. Therefore, we tested stacking 2 or 3 blocks of DyHead, but both resulted in a decrease in overall recognition performance. As a result, adding DyHead is not suitable for the architecture of our research paper.

3.3 Region Proposal Network

In the architecture of the RPN (Region Proposal Network), we use the Cascade RPN[15]. This algorithm specifically addresses the misalignment issue between anchors

doi:10.6342/NTU202304551

and features that occurs during the iterative process of the RPN. The Cascade RPN utilizes a single anchor and combines anchor-based and anchor-free criteria to determine positive samples. It takes advantage of the benefits of multi-stage fine-tuning while maintaining alignment between features and anchors. To achieve this, the Cascade RPN employs adaptive convolution for fine-tuning the anchors at each stage. The adaptive convolution can be seen as a lightweight ROI Align layer. In summary, the Cascade RPN employs adaptive convolution in stage one to align anchors with features. The aligned anchors are then passed to stage two, where classification is performed to determine whether they belong to the background or foreground. Additionally, regression is applied to further refine the bounding boxes. During Stage 1, the adaptive convolution mechanism adjusts the anchors to align with the corresponding features, ensuring accurate localization. In Stage 2, the aligned anchors are evaluated for classification, distinguishing between background and foreground regions. Simultaneously, regression is utilized to refine the bounding box coordinates of the detected objects, improving their accuracy. The entire architecture of the Cascade RPN can be referred to as Fig 3.3(e).



(a) RPN

(b) Iterative RPN

(c) Iterative RPN+

(d) GA-RPN

(e) Cascade RPN

Fig 3.3 The architecture of different networks. “I”, “H”, “C” and “A” denote input image, network head, classifier and anchor regressor, respectively.[15]

doi:10.6342/NTU202304551

By incorporating these stages and operations, Cascade RPN enhances the alignment between anchors and features, leading to improved detection performance.

3.4 Cascade RCNN

Finally, in the model architecture, we have adopted the Cascade RCNN technique[16]. One common issue with proposals generated by the RPN is that a significant portion may have low quality, making it challenging to use a high threshold detector directly. Cascade R-CNN addresses this problem by employing cascade regression as a resampling mechanism, gradually increasing the IoU value of the proposals at each stage. The Cascade RCNN consists of multiple stages, and at each stage, the proposals undergo a refinement process through cascade regression. The initial proposals from the RPN are used as input in the first stage. The proposals that pass a certain IoU threshold in the previous stage are resampled and refined in the current stage. This resampling mechanism ensures that the proposals adapt to higher threshold requirements in subsequent stages.

In the Cascade RCNN framework, each stage's detector is designed to avoid overfitting and ensure that there are enough samples satisfying the threshold conditions. This is achieved by carefully selecting and resampling the proposals based on their IoU with ground truth during each stage. By doing so, the detectors in each stage are trained on a subset of proposals that are more likely to be high-quality detections. With the cascading structure, the deeper stages of the Cascade RCNN are optimized to handle proposals with higher IoU thresholds. As the proposals progress through the cascade, the detectors in each stage focus on refining the proposals and improving their quality. This enables the model to effectively handle more challenging cases and detect objects with higher accuracy. During inference, although the initial proposals generated by the RPN may have lower quality, the quality improves at each stage of the Cascade RCNN. As the proposals

doi:10.6342/NTU202304551

pass through each stage, they undergo refinement and are filtered based on their quality. This iterative process helps align the proposals with the higher IoU thresholds of subsequent stages. As a result, the mismatch between the proposals and detectors with higher IoU thresholds is mitigated, leading to improved detection performance.

3.5 Normalization

In our experiments, we incorporated Group Normalization[24] and Weight Standardization[25] techniques. Group Normalization is a normalization technique that divides the channels of a feature map into groups and normalizes each group independently. This helps in reducing the dependency on batch sizes and improves the generalization ability of the model. Weight Standardization, on the other hand, is a regularization technique that normalizes the weights of the neural network layer-wise. It helps in stabilizing the weight updates during training and has been shown to improve the generalization performance of deep models. By using these techniques, we aim to enhance the training stability, promote better convergence, and improve the overall performance of the model.

3.6 Sampler

In the original RPN, positive and negative samples are randomly sampled based on their IOU with ground truth boxes for training. This sampling technique helps balance the number of positive and negative samples and guides the model's optimization direction. However, if all anchor-generated candidate boxes are included for training without this technique, it would result in many negative samples, which predominantly represent background regions. This imbalance would lead to negative samples dominating the training loss, most of which are easy to classify, thereby deviating the model's optimization from the desired objective.

doi:10.6342/NTU202304551

However, inspired by the success of the focal loss in single-stage detectors, we attempted to implement a similar technique in the RPN training of the two-stage detector. The focal loss reduces the weight of easy-to-classify samples, allowing the model to focus more on challenging samples during training. It uses an appropriate function to measure the contribution of easy and hard samples to the total loss. By leveraging the advantages of focal loss, we can avoid balanced one-to-one sampling of positive and negative samples in RPN training and we can use all bounding box generated by all anchors to join the training, regardless of whether they are positive samples or negative samples. Experimental results have shown that although this approach may slightly reduce training speed and increase memory consumption, it can further improve the accuracy of the model.

3.7 Loss Functions

In our approach, we made specific choices for the loss functions used in different components of the model. For the classification loss in the RPN (Region Proposal Network), we replaced the commonly used cross-entropy loss with focal loss[17].

$$CE(p_t) = -\log(p_t)\,, p_t = \begin{cases} p, if\ y = 1 \\ 1 - p, otherwise \end{cases} \qquad (3-1)$$

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma}\log(p_t)\,, p_t = \begin{cases} p, if\ y = 1 \\ 1 - p, otherwise \end{cases} \qquad (3-2)$$

(3-1) is the cross-entropy loss in binary classification, p is the probability when y=1. (3-2) formula is the focal loss in binary classification, p is the probability when y=1.

The formula (3-1) and (3-2) are compared, where the focal loss has an additional modulating factor $(1 - p_t)^{\gamma}$. For accurately classified samples, p_t approaches 1, causing the modulating factor to approach 0. Conversely, for inaccurately classified samples, the modulating factor approaches 1. As a result, compared to the cross-entropy loss, the focal loss has no significant impact on inaccurately classified samples, but it reduces the loss for accurately classified samples. Overall, this effectively reduces the weight of accurately classified samples in the total loss calculation.

This change was motivated by the sampling strategy mentioned earlier. Focal loss helps address the issue of class imbalance by down weighting easily classified samples and focusing more on hard examples, thereby improving the model's ability to handle challenging cases.

$$L_{ciou} = 1 - IOU + \frac{\rho^2(b, b_t)}{c^2} + v,\ \ v = \frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right) \qquad (3-3)$$

For the bounding box regression loss of the candidate boxes, we introduced the CIOU (Complete Intersection over Union) loss[21] as a replacement for the smooth L1 loss used in the Mask R-CNN paper. Equation (3-3) describes the CIOU (Complete Intersection over Union) Loss. In equation (3-3), $\rho^2(b, b_t)$ represents the Euclidean distance between the center points of the prediction box and the ground truth, while the term v accounts for the aspect ratio difference between the prediction box and the ground truth. Minimizing CIOU involves not only considering the Intersection over Union (IOU) but also ensuring that the prediction box and ground truth center points overlap as much as possible, while their aspect ratios remain similar. We found that CIOU loss leads to faster convergence and achieves better results. CIOU loss considers both the spatial overlap and geometric properties of the bounding boxes, providing a more comprehensive measure for localization accuracy. In the ROI (Region of Interest) head, we followed a similar approach to the cascade R-CNN paper. We used cross-entropy loss for classification and continued to employ the CIOU loss for bounding box regression. This choice aligns with the cascade R-CNN framework and helps maintain consistency throughout the model.

3.8 Data Augmentations

In the study, we encountered the challenge of limited pedestrian recognition datasets, such as CityPersons, which contains only 2,975 images. Without data augmentation, the model would be prone to overfitting due to the limited dataset size. To address this issue, we employed various data augmentation techniques.

The data augmentation methods used in this research include random resizing followed by random cropping. Additionally, we applied random transformations to the images, such as altering the HUE, saturation, brightness, and contrast, performing equalization, and introducing shadows and sun flares (each augmentation with an independent event probability of 0.5). We found that these augmentation techniques significantly improved the issue of overfitting, thereby enhancing the model's generalization capability.

By applying random resizing and cropping, the model learns to adapt to different scales and perspectives of pedestrians in various images. The random transformations, such as altering HUE, saturation, brightness and contrast, introduce variations in color, further diversifying the dataset and reducing the risk of overfitting to specific color distributions. The equalization technique helps to normalize the image histogram, enhancing the visibility of important pedestrian features. Finally, the introduction of shadows and sun flares adds variations in lighting conditions, enabling the model to handle different illumination scenarios.

doi:10.6342/NTU202304551

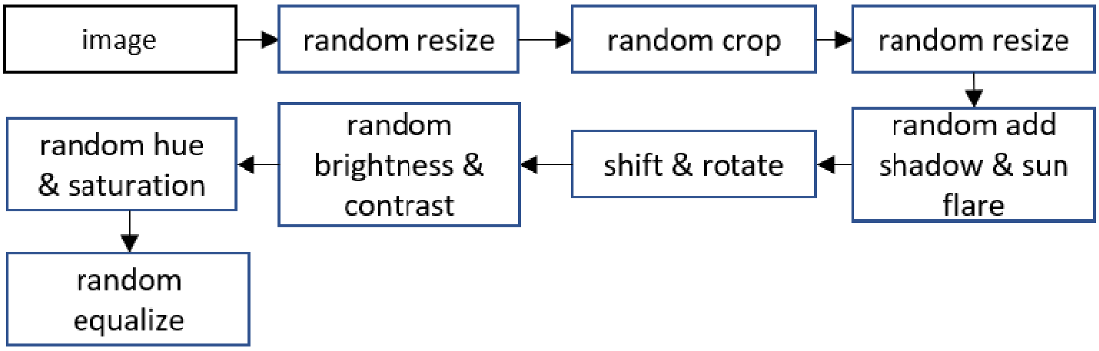


Fig 3.4 data augmentation process

Overall, the combination of these data augmentation techniques effectively mitigates overfitting, enhances the model's ability to generalize to unseen data, and improves pedestrian recognition performance.

3.9 Post Processing

In our post-processing stage, we employed an improved version of NMS called Soft-NMS[29], which addresses the limitations of traditional NMS used in the Faster R-CNN paper. Traditional NMS aggressively eliminates bounding boxes with IOU values greater than the NMS threshold by setting their scores to zero, effectively discarding them. This can result in missed detections, especially in scenes with occlusions.

The Soft-NMS approach takes a different strategy by applying a score penalty mechanism to bounding boxes with IOU values above the threshold. Instead of completely discarding them, the scores of these bounding boxes are reduced using a penalty function that is positively correlated with IOU. When neighboring detection boxes have high IOU values with the current box, they should be suppressed, leading to lower scores. Bounding boxes that are far away from each other remain unaffected. Soft-NMS effectively mitigates the issue of missed detections caused by the aggressive elimination of bounding boxes in traditional NMS, particularly in scenarios involving

doi:10.6342/NTU202304551

object occlusions.

In the field of pedestrian recognition, scenes often involve high occlusion, where multiple pedestrians can be densely packed in specific areas of an image. This poses a challenge for object detection algorithms, as traditional NMS may mistakenly discard bounding boxes due to their overlapping nature.

By using Soft-NMS, we can reduce the issue of false deletion caused by overlapping detection boxes. Soft-NMS applies a penalty mechanism to bounding boxes with high IoU values, effectively suppressing their scores while still retaining them for further consideration. This means that even in densely crowded scenes with overlapping pedestrians, the algorithm can retain and accurately detect the individual pedestrians without discarding their bounding boxes based solely on overlap.

Soft-NMS helps address the specific challenges posed by occlusion in pedestrian recognition scenarios. By considering the degree of overlap and applying a penalty rather than a binary elimination, Soft-NMS improves the overall detection accuracy and ensures that important pedestrian instances are not overlooked.

doi:10.6342/NTU202304551

Chapter 4 Experiments

4.1 Datasets

We conducted experiments using two of the most well-known datasets in pedestrian detection: CityPersons[30] and the Euro City Persons[31] dataset.

CityPersons is a subset of the Cityscape dataset, consisting of 2,975 training images and 19,238 pedestrian instances. The images were collected from 27 major cities in Germany over several months, covering the seasons of spring, summer, and autumn. The images were captured during daylight hours under favorable weather conditions. This dataset offers a wide diversity of scenes due to its varied locations and extended time span. However, a limitation of CityPersons is the relatively small number of training samples available.

The Euro City Persons dataset, on the other hand, provides a larger training set with 21,795 images and 201,323 pedestrian instances. The dataset encompasses 12 different countries and 31 cities, offering a diverse range of scenes throughout all seasons of the year. While the Euro City Persons dataset contains images captured during both daytime and nighttime, for the sake of fair comparison with other methods, we solely used the daytime images in this research.

Both datasets also include various other objects, such as groups of people, motorcyclists, and vehicles, besides pedestrians. However, since the focus of this study is on accurately detecting pedestrians, during training, we considered all other objects as background and solely focused on the task of identifying pedestrians.

By utilizing these datasets, we aimed to evaluate the performance of our model in accurately detecting pedestrians, considering the challenges posed by different lighting conditions, occlusions, and diverse scenes. A detailed comparison of these two datasets is

doi:10.6342/NTU202304551

provided in Table 4.1.

Due to the submission limit imposed by the dataset's test server, we were unable to utilize the testing set for evaluating the model's performance. As a common practice followed in other researches, we used the validation set as a substitute for testing and reported the results based on its evaluation.

Table 4.1 Pedestrian detection datasets summary

datasets	CityPersons	Euro City Persons
images	2975	21795
pedestrians	19238	201323
pedestrians per images	6.47	9.2
cities	27	31
weather	dry	dry, wet
resolutions	2048×1024	1920×1024

4.2 Evaluation metric

To measure the performance of the detector, we utilized the LAMR[32] (Log-average miss rate), also known as MR^{-2}.

To understand the meaning of this metric, it is necessary to first grasp the concepts of MR (miss rate) and FPPI (false positives per image).

$$\text{mr} = \frac{FN(c)}{TP(c) + FN(c)} \qquad (4-1)$$

$$\text{fppi} = \frac{FP(c)}{\#img} \qquad (4-2)$$

MR (miss rate) represents the ratio of missed detections (i.e., the number of false negatives) to the total number of ground truth instances in the dataset. It indicates the percentage of instances that were not detected by the model. FPPI (false positives per image) measures the average number of false positive detections per image. It quantifies the model's tendency to generate false alarms or incorrectly detect objects that are not present.

The evaluation is performed by considering detections with a confidence value equal to or greater than a specified threshold, denoted as **c**. This threshold serves as a control variable commonly used in object detection evaluation. By lowering the threshold value **c**, more detections are included in the evaluation, resulting in a higher number of potential true or false positives and fewer false negatives. And then we can define the formula

$$\mathrm{LAMR} = \exp\left(\frac{1}{9}\sum_{f}\log\left(mr\left(\underset{fppi(c)\leq f}{\operatorname{argmax}}\ fppi(c)\right)\right)\right) \qquad (4-3)$$

The (4-3) formulas can be visualized as representing a two-dimensional coordinate system with MR and FPPI as the vertical and horizontal axes, respectively. By selecting different values for **c**, a curve can be plotted in this coordinate system. The LAMR metric calculates the average of the MR values corresponding to nine logarithmically spaced FPPI values on this curve.

In general, due to the non-rigid nature of pedestrians, we consider any predicted bounding box with an Intersection over Union (IoU) value greater than or equal to 0.5 with the ground truth as a true positive. It means that the predicted bounding box has a significant overlap with the actual pedestrian.

Furthermore, each ground truth can only be matched to one predicted bounding box. If multiple predicted bounding boxes correspond to the same ground truth, we select the one with the highest score and consider it as a true positive, while the other corresponding detections are treated as false positives. After matching all predicted bounding boxes to the ground truth, any ground truth that remains unmatched is considered a false negative. These are the instances where the detector failed to detect the presence of pedestrians.

This evaluation approach ensures that we focus on correctly identifying pedestrians with a reasonable overlap and penalize false positives and false negatives accordingly.

The evaluation of pedestrian detection models considers various criteria to accommodate the variations in size and visibility of pedestrians. These criteria are detailed in Table 4.2 and follow the established conventions used in other references. Differentiating the evaluation based on these criteria allows us to assess the detection performance of the model across different scales and levels of occlusion.

Table 4.2 Evaluation settings for pedestrian datasets based on height and visibility.

Settings	CityPersons		Euro City Persons	
	Visibility	Height	Visibility	Height
Reasonable	[0.65, ∞]	[50, ∞]	[0.6, ∞]	[40, ∞]
Small	[0.65, ∞]	[50, 75]	[0.6, ∞]	[30, 60]
Heavy Occluded	[0.2, 0.65]	[50, ∞]	[0.2, 0.6]	[40, ∞]
All	[0.2, ∞]	[20, ∞]	[0.2, ∞]	[20, ∞]

4.3 Training Settings

We utilized a hardware configuration consisting of a single Nvidia GeForce RTX 4090 for all training and testing processes. Due to budget and cost constraints, we were unable to utilize workstation-grade GPUs such as A100 or V100 for computation. Therefore, within the limitations of our budget, we opted to use the best GPU available for personal

doi:10.6342/NTU202304551

computers, which is the Nvidia GeForce RTX 4090, to conduct our experiments.

Because of the limited VRAM of the Nvidia GeForce RTX 4090, we had to take certain measures during training. Firstly, we applied data augmentation techniques to resize the images to smaller resolutions. Additionally, we had to limit the batch size in our experiments, with the maximum being a batch containing 5 images. These steps were taken to optimize memory usage and ensure efficient training within the constraints of the available hardware resources.

In terms of model architecture parameters, we used the Swin Transformer with a base size, a drop rate of 0.2, and pretraining on the ImageNet-22k dataset at a resolution of 224 x 224 pixels. For detailed information on the model size settings, you can refer to Table 4.3

Table 4.3 Detailed architecture of the Swin Transformer

	Swin-Tiny	Swin-Small	Swin-Base	Swin-Large
stage 1	[dim 96, head 3] × 2	[dim 96, head 3] × 2	[dim 128, head 4] × 2	[dim 192, head 6] × 2
stage 2	[dim 192, head 6] × 2	[dim 192, head 6] × 2	[dim 256, head 8] × 2	[dim 384, head 12] × 2
stage 3	[dim 384, head 12] × 6	[dim 384, head 12] × 18	[dim 512, head 16] × 18	[dim 768, head 24] × 18
stage 4	[dim 768, head 24] × 2	[dim 768, head 24] × 2	[dim 1024, head 32] × 2	[dim 1536, head 48] × 2
window size for all stage	7 × 7	7 × 7	7 × 7	7 × 7

In our configuration of PAFPN, we consistently output a 5-level FPN pyramid, with each level having a channel size of 256. This choice of the number of levels and channel

doi:10.6342/NTU202304551

size aims to strike a balance between capturing multi-scale information and maintaining computational efficiency. By having a 5-level FPN pyramid, we can effectively capture features at different scales, enabling the model to handle objects of various sizes. The choice of a channel size of 256 allows for sufficient expressive power while keeping the computational demands manageable.

In the ROI (Region of Interest) align layer, we observed that increasing the size of the ROI feature map from 7×7 to 14×14 improves the precision of bounding box localization. By increasing the size of the ROI feature map, we provide a finer level of spatial resolution, allowing for more precise localization of the objects within the region proposals. The larger feature map size enables the model to capture more detailed information about the objects, which helps improve the accuracy of bounding box regression. By adjusting the ROI feature map size to 14x14, we strike a balance between computational efficiency and localization accuracy, achieving better results in this task.

Finally, we utilized the AdamW optimizer for training the model. We set the initial learning rate to 0.001 and incorporated a warm-up phase for the first 1000 iterations, gradually increasing the learning rate to its initial value. After the 20th epoch, we decayed the learning rate to 0.0001 for further fine-tuning. The total training duration consisted of 40 epochs.

4.4 Results

In this section, we present a comprehensive comparison between our proposed method and other state-of-the-art approaches in the field. The comparison is based on the evaluation metric LAMR, also known as MR^{-2}, which was discussed earlier. The detailed comparison results are presented in Table 4.4

doi:10.6342/NTU202304551

Table 4.4 Comparison of our method with the current state-of-the-art detectors based on MR^{-2}

Method /MR^{-2}	Reasonable	Small	Heavy Occ	All
City Persons				
MS-CNN[33]	13.32%	15.56%	51.88%	39.94%
Cascade R-CNN[34]	11.62%	13.64%	47.14%	37.63%
Repulsion Loss[35]	11.48%	15.67%	52.59%	39.17%
AdaptiveNMS[36]	11.40%	13.64%	46.99%	38.89%
OR-CNN[37]	11.32%	14.19%	51.43%	40.19%
HBA-RCNN[38]	11.26%	15.68%	39.54%	38.77%
DVRNet[39]	11.17%	15.62%	42.52%	40.99%
MGAN[40]	9.29%	11.23%	40.97%	38.86%
Ours	**11.46%**	**15.10%**	**37.63%**	**34.76%**
Euro City Persons				
SSD[41]	10.5%	20.5%	42.0%	-
YOLOv3[41]	8.5%	17.8%	37.0%	-
Faster R-CNN[41]	7.3%	16.6%	52.0%	-
Cascade-RCNN[34]	6.6%	13.6%	33.3%	-
F2D-Net[42]	6.1%	10.7%	28.2%	-
Ours	**5.7%**	**13.8%**	**28.0%**	**18.4%**

The Table 4.4 provided in the previous section demonstrates a fair and just comparison between our model and other state-of-the-art models. All models listed in the table were trained solely on their respective datasets, without any additional training data from other sources. For instance, the models evaluated on the Euro City Persons dataset were only trained on the Euro City Persons training dataset, and they did not utilize any data from the CityPersons or other pedestrian datasets, such as Caltech. Similarly, our model was trained exclusively on the Euro City Persons training data without incorporating data from other sources. By ensuring that all compared models adhere to this criterion, we maintain fairness and consistency in the evaluation process. This approach guarantees that each model's performance is solely based on the data available within its respective dataset and avoids any potential biases that could arise from using additional training data.

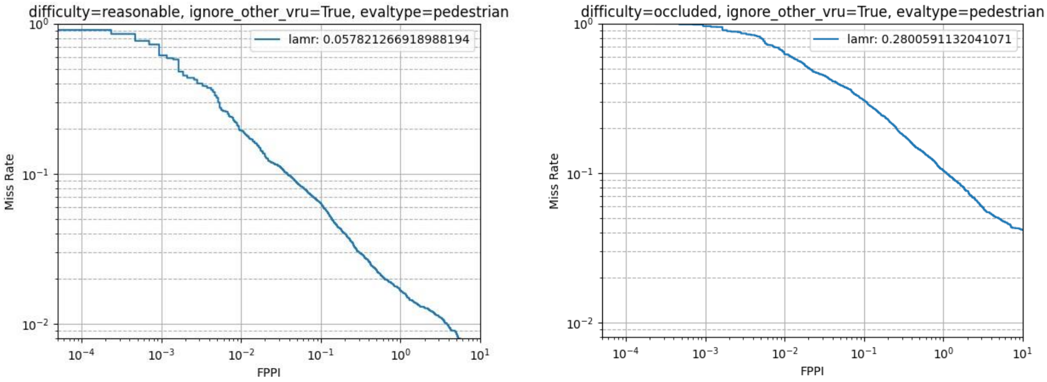


Fig 4.1 LAMR on the Euro City Persons dataset

Fig 4.1 illustrates the LAMR (Log Average Miss Rate) of the experiment on the Euro City Persons dataset. As mentioned earlier when introducing LAMR, as MR (Miss Rate) decreases, FPPI (False Positives Per Image) tends to increase. Therefore, LAMR serves as an indicator that balances both characteristics and impartially assesses the model's performance.

Based on our experimental findings, our model demonstrated outstanding performance on the Euro City Persons dataset, surpassing even CNN-based models in reasonable scenarios. However, its efficacy on the CityPersons dataset was comparatively weaker, showing slightly inferior performance in various situations, except for cases involving occlusion. We attribute this difference in performance to the nature of transformer-based models, which typically require a larger amount of training data for optimal performance.

With the relatively limited training data available in the CityPersons dataset, it is expected that the performance of the transformer-based model may not match that of traditional CNN-based models. Despite the somewhat lower performance on the CityPersons dataset, we found a noteworthy trend in our model's detection capabilities. Specifically, when it comes to heavily occluded pedestrians, our model outperformed many other models. This result strongly indicates that our model is highly suitable for scenarios involving occluded pedestrians. The capability of our model to excel in detecting heavily occluded pedestrians highlights its potential in real-world scenarios where pedestrians are frequently obstructed by various objects or other individuals. While further improvements may be possible for detecting pedestrians in reasonable and small scales, our model's proficiency in handling heavily occluded pedestrians is a promising and valuable aspect of its performance.

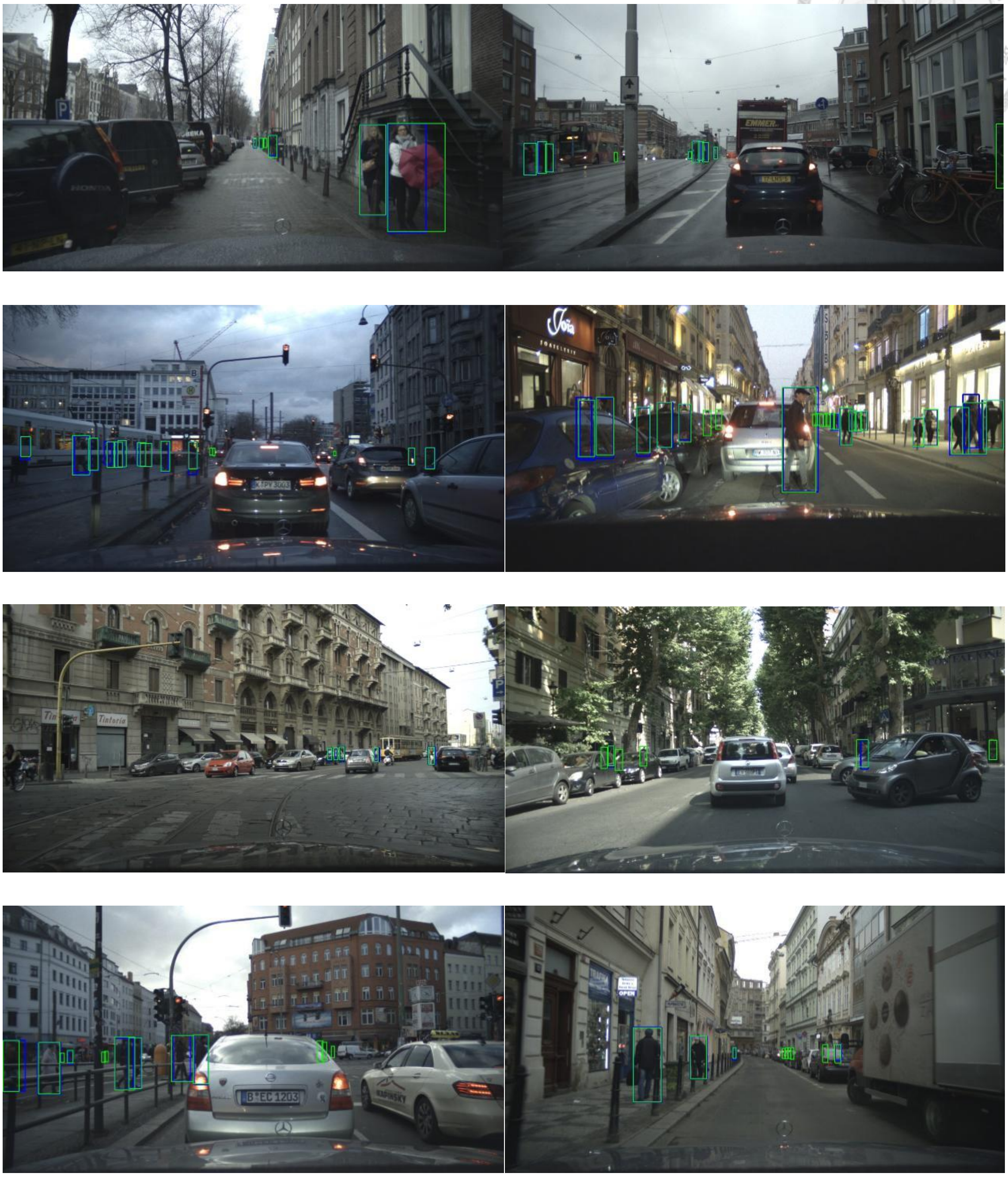


Fig 4.2 Detection result: the green rectangle is our predictions and the blue ones is the ground truths.

doi:10.6342/NTU202304551

Chapter 5 Conclusion

In this study, we proposed a novel pedestrian detection model based on the Swin Transformer architecture. Our model leveraged the powerful capabilities of the Swin Transformers to capture long-range dependencies and spatial relationships in images, making it well-suited for the challenging task of pedestrian detection. Through a series of carefully designed modules and techniques, we aimed to address various issues, including occlusions and scale variations, commonly encountered in pedestrian recognition scenarios. We also employed data augmentation techniques, including random resizing, cropping, and various image transformations, to mitigate overfitting due to the limited size of the training datasets. Additionally, we used soft-NMS for post-processing, which improved the detection results, especially in scenarios with high levels of occlusion. For loss functions, we employed focal loss in the RPN stage and CIOU loss for bounding box regression, which enhanced the model's convergence speed and detection accuracy. Furthermore, we incorporated group normalization and weight standardization, contributing to stable and efficient training. Our experiments on the Euro City Persons and CityPersons datasets demonstrated promising results. In particular, our model exhibited exceptional performance in detecting heavily occluded pedestrians, showcasing its potential for handling challenging scenarios where traditional methods may struggle. However, we also acknowledged that the performance on CityPersons was not as strong as expected, which we attributed to the limited amount of training data available for a Transformer-based backbone. This finding highlighted the need for larger datasets to fully unleash the capabilities of our model. Our proposed pedestrian detection model based on the Swin Transformer architecture demonstrated promising results, particularly in scenarios with heavily occluded pedestrians. We showcased its potential to handle

doi:10.6342/NTU202304551

challenging detection tasks and provided fair comparisons with other state-of-the-art models to objectively assess its performance. However, further improvements could be achieved with larger training datasets and more powerful hardware. Our research contributes valuable insights and sets a foundation for future advancements in pedestrian detection using Transformer-based models.

doi:10.6342/NTU202304551

REFERENCE

[1] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 2016: Springer, pp. 21-37.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems,* vol. 28, 2015.

[3] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117-2125.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961-2969.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693-5703.

[8] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 2020: Springer, pp. 213-229.

[10] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012-10022.

[11] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759-8768.

[12] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7036-7045.

[13] K. Chen, Y. Cao, C. C. Loy, D. Lin, and C. Feichtenhofer, "Feature pyramid grids," *arXiv preprint arXiv:2004.03580,* 2020.

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, 2001, vol. 1: Ieee, pp. I-I.

[15] T. Vu, H. Jang, T. X. Pham, and C. Yoo, "Cascade rpn: Delving into high-quality region proposal network with adaptive convolution," *Advances in neural information processing systems,* vol. 32, 2019.

[16] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154-6162.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object

doi:10.6342/NTU202304551

detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980-2988.

[18] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 516-520.

[19] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658-666.

[20] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, no. 07, pp. 12993-13000.

[21] Z. Zheng *et al.*, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Transactions on Cybernetics,* vol. 52, no. 8, pp. 8574-8586, 2021.

[22] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing,* vol. 506, pp. 146-157, 2022.

[23] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 761-769.

[24] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3-19.

[25] S. Qiao, H. Wang, C. Liu, W. Shen, and A. Yuille, "Micro-batch training with batch-channel normalization and weight standardization," *arXiv preprint arXiv:1903.10520,* 2019.

[26] Z. Liu *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12009-12019.

[27] Y. Zheng, X. Zhang, R. Zhang, and D. Wang, "Gated Path Aggregation Feature Pyramid Network for Object Detection in Remote Sensing Images," *Remote Sensing,* vol. 14, no. 18, p. 4614, 2022.

[28] X. Dai *et al.*, "Dynamic head: Unifying object detection heads with attentions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7373-7382.

[29] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS--improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561-5569.

[30] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3213-3221.

[31] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "The eurocity persons dataset: A novel benchmark for object detection," *arXiv preprint arXiv:1805.07193,* 2018.

[32] 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, 2011.

[33] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October*

doi:10.6342/NTU202304551

11–14, 2016, Proceedings, Part IV 14, 2016: Springer, pp. 354-370.

[34] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE transactions on pattern analysis and machine intelligence,* vol. 43, no. 5, pp. 1483-1498, 2019.

[35] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7774-7783.

[36] S. Liu, D. Huang, and Y. Wang, "Adaptive nms: Refining pedestrian detection in a crowd," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6459-6468.

[37] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 637-653.

[38] R. Lu, H. Ma, and Y. Wang, "Semantic head enhanced pedestrian detection in a crowd," *Neurocomputing,* vol. 400, pp. 343-351, 2020.

[39] L. Shi, C. Livermore, and I. A. Kakadiaris, "DVRNet: Decoupled Visible Region Network for Pedestrian Detection," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 2020: IEEE, pp. 1-9.

[40] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network for occluded pedestrian detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4967-4975.

[41] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "Eurocity persons: A novel benchmark for person detection in traffic scenes," *IEEE transactions on pattern analysis and machine intelligence,* vol. 41, no. 8, pp. 1844-1861, 2019.

[42] A. H. Khan, M. Munir, L. van Elst, and A. Dengel, "F2DNet: fast focal detection network for pedestrian detection," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022: IEEE, pp. 4658-4664.

doi:10.6342/NTU202304551