ENHANCING PEDESTRIAN TARGET RECOGNITION IN OPEN COMMUNITY MULTI-SCENE SPACES USING THE YOLO-STP NETWORK

Chun Liu[1], Yanyi Li[1,*], Jiajing Gu[1], Yongqi Lou [2], Tao Shen[2]

[1]College of Surveying and Geo-informatics, Tongji University, Shanghai, China – (liuchun, liyanyi19981104, gujiajing) @tongji.edu.cn
[2] College of Design and Innovation, Tongji University, Shanghai, China – (louyongqi, shentao) @tongji.edu.cn

KEY WORDS: Open community, Pedestrians, NICE2035, Deep learning.

ABSTRACT:

Addressing the challenge of quantitatively analyzing and presenting pedestrian elements within open community spaces is of significant importance. Focusing on the indoor scene spaces of open communities, this study introduces the TJ-Person pedestrian target recognition image dataset. Furthermore, we design a deep learning-based community pedestrian activity analysis network model and incorporate various attention mechanisms, such as SA, CA, CBAM, SE, and SK, into the YOLO v5s deep learning target recognition network framework for comparative evaluation of pedestrian target recognition in open communities. Utilizing the optimized YOLO Swin Transformer Person (YOLO-STP) network, precise identification of pedestrian targets across multiple scenarios was achieved. We conducted experimental verification using four typical scenarios within Shanghai's NICE2035 open community as case studies. The results demonstrated that the proposed YOLO-STP community pedestrian activity analysis network model achieved an optimal detection accuracy of up to 98.47%. In all four tested scenarios, the YOLO-STP method consistently exhibited competitive performance. Moreover, in the COCO-2017 open-source dataset testing, the YOLO-STP method outperformed other networks of the same type, showcasing its significant advantages. Overall, the research presented in this study provides a crucial technical foundation for the analysis and recognition of pedestrian targets in future community scenarios.

1. INTRODUCTION

As urbanization progresses rapidly, community spatial structures are becoming increasingly complex, with open communities emerging as the primary residential areas (Carmona et al., 2010). Within these communities, pedestrian target recognition technology holds significant application value in safety management, indoor navigation, and public activity organization (Dollár et al., 2012). Nevertheless, the complexity and diversity of indoor environments present numerous challenges for traditional pedestrian target recognition methods when applied to the multifaceted scene spaces of open communities. To address these challenges, researchers have begun exploring the use of deep learning networks for pedestrian target recognition (Zhao et al., 2005; Zhang et al., 2022).

Deep learning networks have achieved substantial advancements in the field of computer vision, particularly in object detection and recognition (Liu et al., 2016). The YOLO series networks, for instance, have garnered widespread attention owing to their end-to-end real-time performance and high recognition accuracy (Redmon et al., 2016). Despite the YOLO networks' remarkable performance in various scenarios, there is still room for improvement concerning pedestrian target recognition in multi-scene spaces of open communities (Diraco et al., 2015). Consequently, researchers have started exploring the integration of attention mechanisms to enhance the recognition accuracy of YOLO networks in such contexts (Woo et al., 2018).

Analyzing the activity characteristics and patterns of pedestrians within a scene space holds immense significance for subsequent research on the vitality of community spaces. To achieve this objective, establishing a robust foundation for precise recognition and extraction of pedestrian targets becomes essential. Accordingly, this article aims to accurately identify pedestrian targets, which in turn will serve as a crucial technical basis for subsequent analysis of community spatial vitality. The investigation and analysis of pedestrians play a crucial role in advancing the quantification of pedestrian elements within a scene, thereby establishing a solid research foundation for incorporating pedestrian attributes into a wider range of geographic information scene models.

This study presents an improved YOLO-STP deep learning network method for pedestrian target recognition in multi-scene spaces of open communities. This method incorporates various attention mechanisms into the YOLO v5s network, such as SA, CA, CBAM, SE, and SK. The effectiveness of this approach is verified through experiments conducted in four typical scenarios—bar, kitchen, activity room, and dining area—within the NICE2035 open community in Shanghai. Our research demonstrates that the proposed YOLO-STP community pedestrian activity analysis network model achieves an optimal detection accuracy of up to 98.47%. In testing on the COCO-2017 open-source dataset, the YOLO-STP method exhibits significant advantages compared to other networks of the same type.

2. METHODOLOGY

2.1 Basic Structure of YOLO v5 Deep Learning Network

The fundamental network structure employed in this study is based on the YOLO v5 architecture, with the schematic diagram of the network structure illustrated in Figure 1. The network primarily comprises four components: Input, Backbone, Neck, and Head.

[*] Corresponding author

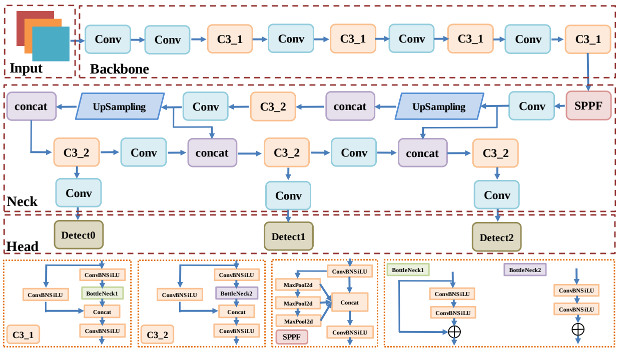


Figure 1. Schematic diagram of YOLO v5 network structure.

Compared to the traditional YOLO series network (Diwan et al., 2023), the YOLO v5 network offers advantages across the four main components. Firstly, at the input stage, several optimization points are introduced: (1) Mosaic data augmentation is used for random scaling, cropping, and layout operations on small targets. (2) Adaptive anchor box calculation is incorporated, initially setting the length and width dimensions of the anchor box, outputting a prediction box, comparing it with the actual box, and updating and optimizing the parameters in reverse. (3) The minimal amount of black edges is added to the original image through adaptive image scaling, calculating the scaling ratio and scaled size, and further determining the value of black edge background filling.

Secondly, in the Backbone section, two optimization points are introduced: (1) Slicing operations are conducted using the Focus structure, and the YOLO v5s network selected in this study employs 32 convolutional kernels. (2) Two CSP structures are defined, corresponding to (C3_1) and (C3_2) in Figure 3. The first CSP structure is applied to the Backbone section, while the second CSP structure is utilized in the Neck section.

Lastly, at the output stage, CIOU_Loss is employed as a loss function, and DIOU_NMS non-maximum suppression is also used, significantly improving the detection of overlapping and occluded targets.

2.2 Principles of Multiple Attention Mechanism Modules

In order to optimize the traditional YOLO v5 network structure, this study incorporates the attention mechanism module. Here, we briefly introduce the basic principles of five attention mechanism modules: SA (Shuffle Attention), CA (Channel Attention), CBAM (Convolutional Block Attention Module), SEAttention (SENet Attention), and SKAttention (Selective Kernel Attention).

2.2.1 ShuffleAttention Mechanism Module (SA)

The design concept of the ShuffleAttention attention mechanism module integrates group convolution, spatial attention mechanism, and channel attention mechanism, utilizing Channel Shuffle to fuse information between different groups. The network structure of this module is depicted in Figure 2 (Zhang et al., 2021).

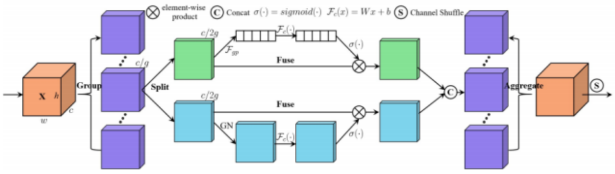


Figure 2. Schematic diagram of SA attention mechanism module.

As illustrated in Figure 3, the tensor is initially divided into g groups, with each group internally processed using the SA Unit. SA is further divided into spatial attention mechanisms, as demonstrated in the blue section. This specific implementation employs GroupNorm (GN) to obtain spatial dimension information. The channel attention mechanism utilized internally by SA is depicted in the green section, with its specific implementation resembling the SE attention mechanism module described below. The SA Unit integrates information within the group through Concate. Lastly, the Channel Shuffle operation is used to rearrange the groups, enabling information flow between different groups.

2.2.2 Coordinate Attention Mechanism Module (CA)

The CA attention mechanism module aims to enhance the expression ability of learning features in mobile networks. It can transform and change any intermediate feature tensor $X = [x_1, x_2, \dots, x_c] \in R^{H\times W\times C}$ in the network and output tensor $Y = [y_1, y_2, \dots, y_c] \in R^{H\times W\times C}$ of the same size. The implementation process of CA attention mechanism is shown in Figure 3 (Hou et al., 2021).

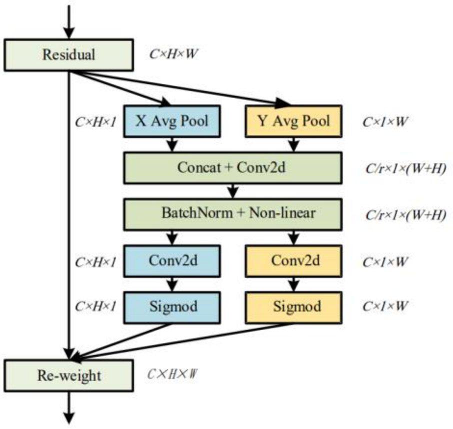


Figure 3. Schematic diagram of CA attention mechanism module.

To acquire attention on image width and height while encoding precise position information, CA initially divides the input feature map into two directions: width and height. Global average pooling is then performed to obtain feature maps in both directions, as illustrated in formulas (1) and (2).

$$Z_c^h(h) = \frac{1}{W}\sum_{0\le i\le W} |x_c(h,i) \qquad (1)$$

$$Z_c^w(w) = \frac{1}{H}\sum_{0\le j\le H} |x_c(j,w) \qquad (2)$$

Then, the feature maps of the width and height of the global Receptive field are spliced together, and then they are sent to the shared convolution core as the convolution module of 1×1, whose dimension is reduced to the original C/r, and then the feature map F1 after batch normalization is sent to the sigmoid activation function to get the feature map f shaped like 1×(W+H)×C/r, as shown in the formula (3).

$$f = \delta(F_1([z^h, z^w])) \qquad (3)$$

Then, the feature map f is convolved into 1×1 according to the original height and width to obtain the feature map F_h and F_w with the same channel number as the original. After the sigmoid activation function, the attention weight g_h in the height and width of the feature map and the attention weight g_w in the width direction are obtained respectively. As shown in formulas (4) and (5).

$$g^h = \sigma(F_h(f^h)) \tag{4}$$

$$g^w = \sigma(F_w(f^w)) \tag{5}$$

After the above calculation, the attention weight g^h in the height direction and the attention weight g^w in the width direction of the input feature map will be obtained. Finally, by multiplying and weighting the original feature map, the final feature map with attention weights in the width and height directions will be obtained. The formula is shown in (6).

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{6}$$

2.2.3 CBAM Attention Mechanism Module (CBAM)

CBAM (Woo et al., 2018) and BAM both employ channel attention modules and spatial attention modules. However, the distinction between them lies in the arrangement of these modules: BAM's channel attention and spatial attention modules are parallel, whereas CBAM connects them in series. CBAM first processes the input features using a channel attention module, followed by a spatial attention module. The overall structure is illustrated in Figure 4.

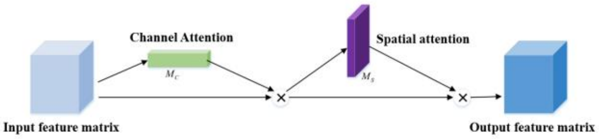


Figure 4. Schematic diagram of CBAM attention mechanism module.

From the overall structure of CBAM, it can be seen that if the input feature is F, the feature map of channel attention is M_c, the output feature of channel attention in CBAM is F’, the spatial attention feature map is M_s, and the output through the spatial attention module is F’’, the mathematical expression of CBAM is shown in formula (7) (8).

$$F' = M_C(F) \tag{7}$$

$$F'' = M_s(F) \tag{8}$$

To extract more channel feature information, the channel attention module of CBAM incorporates maximum pooling and average pooling layers before the MLP (Multi-Layer Perceptron). Simultaneously, the use of two pooling layers enables the extraction of more refined feature information, enhancing the network's expressiveness. The channel attention module is depicted in Figure 5.

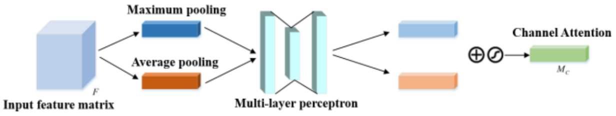


Figure 5. Channel attention mechanism of CBAM.

The mathematical expression of the channel attention module is shown in (9).

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{9}$$

It can be seen from the mathematical expression and figure that the input feature F is input into a shared multi-layer perceptron after being maximally pooled and averagely pooled respectively, and then the two features output from the multi-layer perceptron network are superimposed and input into the activation function to finally obtain F′.

The spatial attention module of CBAM also incorporates maximum pooling and average pooling, resulting in three fewer convolutions compared to BAM. The overall structure is illustrated in Figure 6.

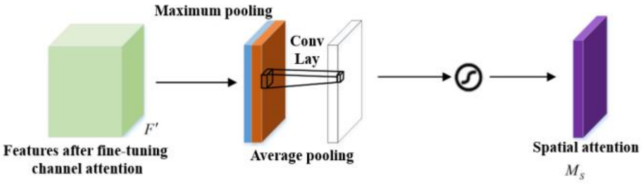


Figure 6. Spatial attention mechanism of CBAM.

It can be seen from Figure 7 that after the input F has passed the maximum pooling and the average pooling, input a 7 × 7 convolution, and finally input it into the activation function. The corresponding mathematical expression is shown in (10).

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \tag{10}$$

2.2.4 SENet Attention Mechanism Module (SE)

SENet (Hu et al., 2019) is a plug-and-play attention module proposed by Hu et al. in 2017, designed to learn specific information in deep networks and generally used after convolutional modules. The overall structure of the SENet module is depicted in the figure. The module primarily performs three operations on the convolutional feature maps: Squeeze, Excitation, and Scale. The overall structure is illustrated in Figure 7.

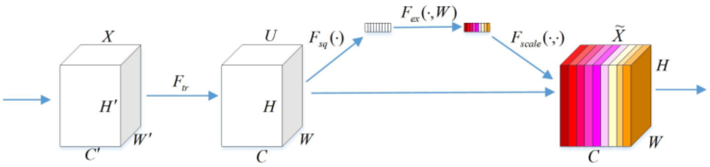


Figure 7. SE module schematic diagram.

In the Squeeze operation, the feature maps are primarily converted into 1×1×C vectors through global pooling, where C represents the number of channels. Through this method, we achieve spatial compression, where the number of bold channels represents the evaluation scores on the corresponding channels. The operation is demonstrated in equation (11).

$$z_c = F_{sq}(u_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{11}$$

In the Excitation operation, to establish a correlation model between channels, two fully connected layers are employed, along with a ReLU activation function to perform nonlinear transformation between channel features. After learning, weights are generated for each feature channel. The formula is illustrated in equation (12).

$$\text{s} = F_{ex}(Z,W) = \sigma\big(g(Z,W)\big) = \sigma(W_2\delta(W_1 Z)) \tag{12}$$

In the Scale operation, the goal is to multiply the channel weights obtained after the Excitation operation with the two-dimensional matrix of the corresponding channel in the original feature map. In other words, weighting the original spatial graph in the channel direction. The formula is illustrated in equation (13).

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c u_c \tag{13}$$

2.2.5 Selective Kernel Attention Mechanism Module (SK)

In standard convolutional neural networks (CNNs), the receptive field of each layer of artificial neurons is designed to share the same size. It is well known in the field of neuroscience that the size of the receptive field of neurons in the visual cortex is regulated by stimulation; however, stimulation is seldom considered when constructing CNNs. This paper introduces a dynamic selection mechanism in CNNs, allowing each neuron to adjust its receptive field size adaptively according to the multi-scale input information. A building block called Selective Kernel (SK) unit is designed, wherein multiple branches of different kernel sizes are fused using softmax attention guided by the information in these branches. Different attention to these branches leads to different sizes of effective receptive fields for neurons in the fusion layer. Multiple SK units are stacked in a deep network called Selective Kernel Networks (SKNets), forming the SKAttention attention mechanism module (Li et al., 2019). The structural diagram of the SKAttention attention mechanism module is shown in Figure 8.

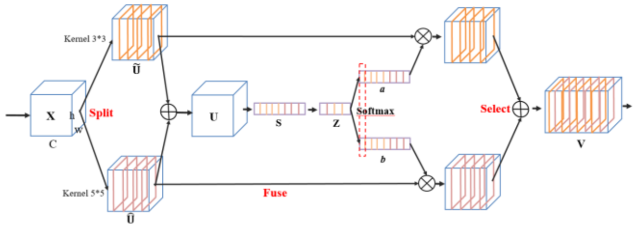


Figure 8. SKAttention module structure diagram.

The main processing process of this module is divided into the following three parts:

(1) Split: Perform complete convolution operations (group convolution) on the input vector X with different kernel sizes. Specifically, in order to further improve efficiency, replace the traditional convolution of 5x5 with a hollow convolution with dimension=2 and a convolution kernel of 3 × 3;

(2) Fuse: After adding two feature maps, perform a global average pooling operation. The fully connected layer that first reduces dimensionality and then increases dimensionality is a two-layer fully connected layer. The output two attention coefficient vectors a and b, where a+b=1;

(3) Select: Select uses two weight matrices, a and b, to weight the previous two feature maps, and there is an operation similar to feature selection between them.

2.3 Construction of YOLO-STP Pedestrian Target Detection Network

In this study, based on the architecture of YOLO v5s, we introduced a multi-channel SKAttention mechanism to optimize the Neck part of the network, a TransformerV2 structure to optimize the Backbone part (Liu et al., 2022), and a Decoupled structure to improve the Head part (Liu et al., 2018). Overall, we formed a YOLO-STP network suitable for multi-scene pedestrian target recognition in open communities. The network structure diagram is shown in Figure 9.

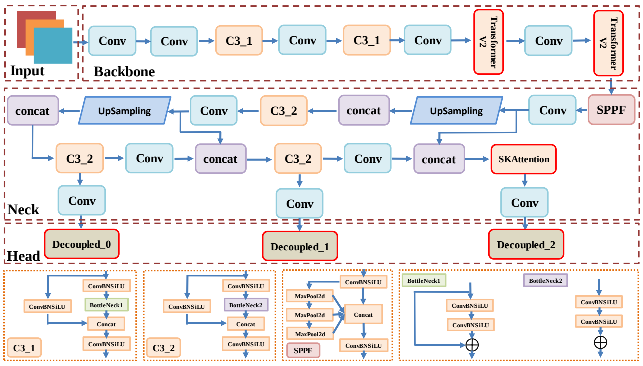


Figure 9. Schematic diagram of YOLO-STP network structure.

The channel attention mechanism used in Figure 9 is the SKAttention module, which was discussed in the previous section and also uses the Decoupled structure. The schematic diagram of this structure is shown in Figure 10.

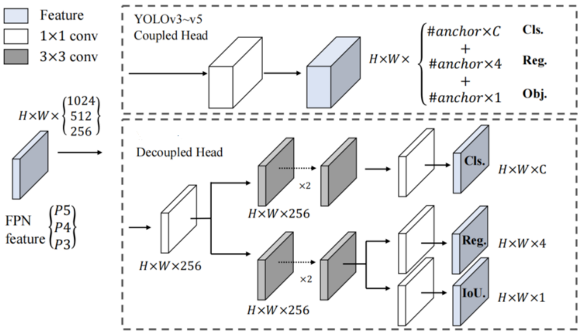


Figure 10. Decoupled structure diagram.

The improvement of the Head section using this structure mainly takes into account the different focuses of classification and localization; therefore, using different branches for operations is beneficial for improving effectiveness. For instance, to avoid a significant increase in computational complexity when using the Decoupled Head structure in Figure 10, for each level of FPN features, we first employ a transformation layer to reduce the feature channel to 256. Subsequently, we add two parallel branches, each with two sets of three transformation layers for classification and regression tasks. Compared to the traditional Head structure of the YOLO network, this approach further enhances the detection effect and speed.

In this study, we innovatively introduce the structural layer of the new version of Transformer V2 to address the issues faced by the traditional Transformer V1 structural layer (Liu et al., 2022). The main improvement is reflected in the red marked part in Figure 11. The traditional Transformer V1 structural layer encounters three main problems: (1) significant training instability may occur when increasing the size of the visual model; (2) in many downstream tasks that require high resolution, there has not been a well-explored method for migrating models trained at low resolution to larger scale models; and (3) in complex background environments, a small number of pixels can cause significant overall interference.

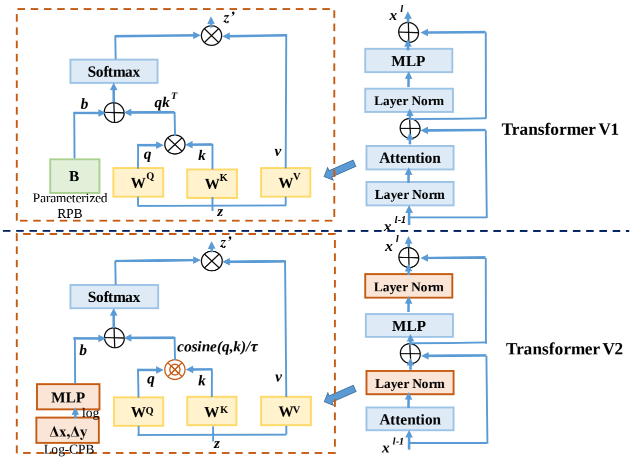


Figure 11. Improvement of V2 relative to V1 version.

To address the first issue, the problem of unstable training, the post-norm approach is adopted, which moves the Layer Norm layer in the Transformer block from the front to the back of the Attention layer. The advantage of doing so is that after calculating the Attention, the output will be normalized, stabilizing the output value. This modification helps to improve the training stability of the model, especially when increasing the size of the visual model.

To address the second issue, this module employs log-space continuous position bias technology to transfer low-resolution pre-trained models to high-resolution models. Instead of using the traditional method, the continuous position offset method is adopted. The principle of this method is shown in equation (14), where a meta-network is used for relative coordinates. This technique enables the smooth migration of models trained at low resolutions to larger-scale models, improving their performance in downstream tasks that require high resolution.

$$B(\Delta x,\Delta y)=G(\Delta x,\Delta y) \tag{14}$$

In equation (14), G represents a small network that generates bias parameters for any relative coordinate, enabling the migration of any variable window size. To address the issue of having a large proportion of relative coordinate ranges that require extrapolation when migrating across large windows, the log space continuous position bias technology is introduced here. Further details regarding this technology are presented in (15).

$$\begin{cases}\widehat{\Delta x}=sign(x)\cdot\log(1+|\Delta x|)\\ \widehat{\Delta y}=sign(y)\cdot\log(1+|\Delta y|)\end{cases} \tag{15}$$

The logarithmic operation is used here because it reduces the extrapolation ratio required for block resolution transfer. This can help in the migration of low-resolution pre-trained models to high-resolution models, making it easier to transfer information across larger window sizes.

To address the third issue, we observed that in the V1 version of self-attention calculation, the dot product of query and key is used to calculate the similarity of pixel pairs. However, in scenarios with a large amount of data, certain modules and heads may be dominated by a small number of pixel pairs in the attention maps. To alleviate this problem, we adopt the Scaled Cosine Attention (SCA) method (Liu et al., 2022), which is shown in formula (16).

$$Sim(q_i,k_i)=cos(q_i,k_i)/\tau+B_{ij} \tag{16}$$

3. EXPERIMENT AND RESULTS

3.1 Data collection and preprocessing

The experiment was conducted at the NICE2035 open community space located near Tongji University in Yangpu District, Shanghai (Lou et al., 2018; Wang et al., 2022; Wang et al., 2021). For this study, we selected four typical scenarios, namely the kitchen, activity room, dining area, and bar scene, to construct and deploy visual sensor networks. These areas were all located in the same spatial and temporal environment, and a unified visual sensor network was used for data collection and time synchronization to ensure that data processing was carried out under a unified time benchmark. Figure 12 (a) shows the four data collection areas, while Figure 13 (b) shows the unified visual sensor network.

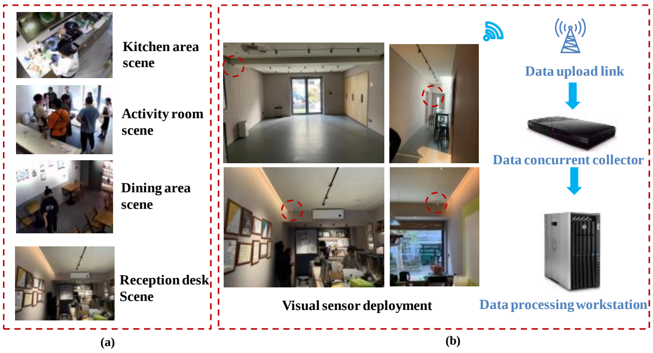


Figure 12. Schematic diagram of experimental scenario selection and data collection.

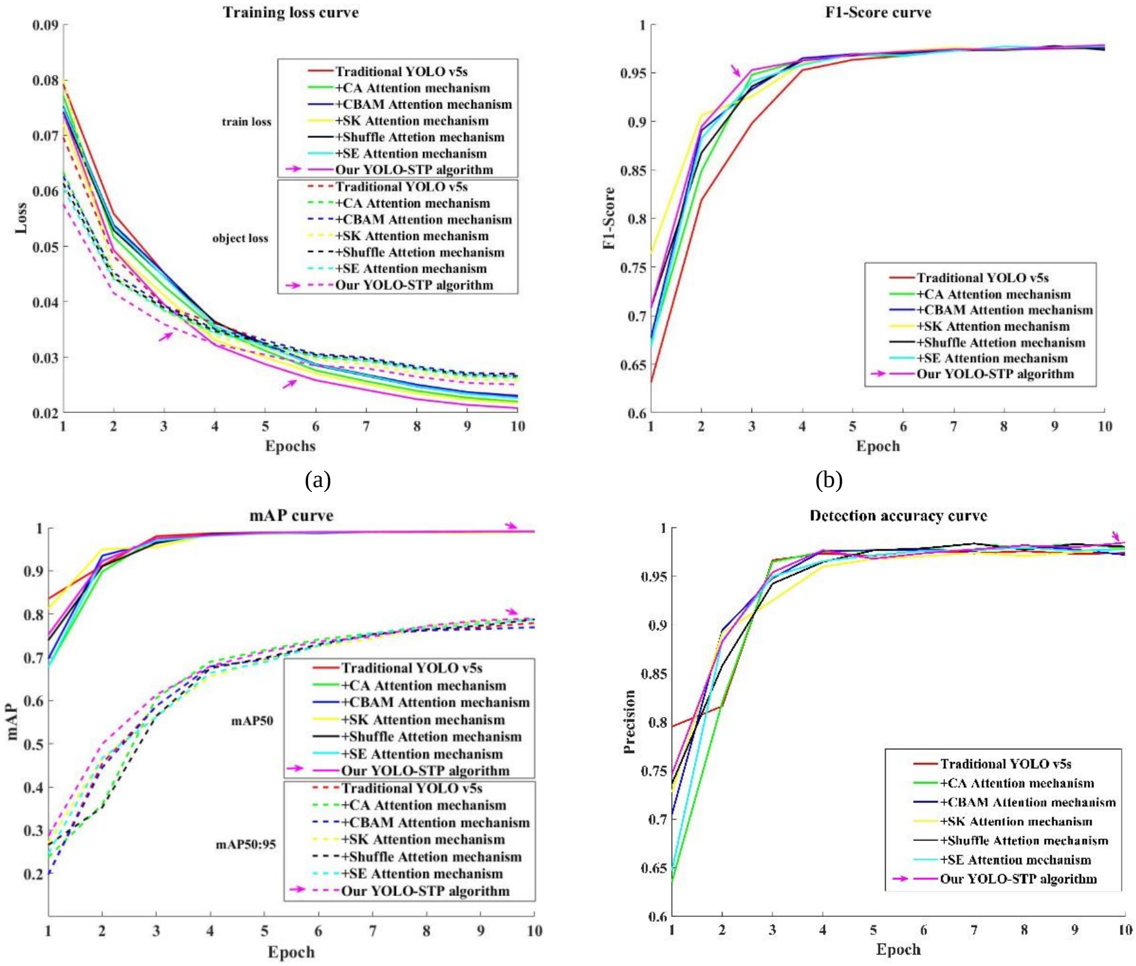
To analyze pedestrian data in the four scenarios, we recorded video data over a 5-day period from December 6 to 10, 2021, corresponding to Monday to Friday, for each scenario. This allowed for an analysis of pedestrian activities in open communities. To train a pedestrian target recognition dataset applicable to multiple scenarios in open communities, we extracted 7002 images with pedestrian targets at a time interval of 100 seconds from the video data collected during the experimental period. The pedestrian targets were labeled through human marking, and we created the TJ-Person dataset for further research, consisting of 6301 images in the training set and 701 images in the validation set. The VPIMT Software was used for data labeling, which can process video data and label pedestrian targets, and can be obtained through the following link: https://github.com/17863958533/VPlabelImg_LYY. The TJ-Person dataset is developed to analyze the activity patterns of individuals in open community spaces by considering the functionality of individuals in space, the continuity of time, and the distinct characteristics of different scenarios. Unlike traditional datasets like ETH Pedestrian, CityPersons, and KITTI, which solely focus on personnel identification and testing, video datasets centered on open community spaces provide an opportunity to delve deeper into the activity patterns of individuals across various scenarios. This expanded scope of data analysis moves beyond mere model effectiveness comparison and holds practical research value.

3.2 Comparison of Multiple Attention Mechanism Improvement Methods

To select the optimal attention mechanism module from the available options, this study conducted experiments by introducing the five attention modules mentioned earlier to the traditional YOLO v5s network architecture for comparison. The initial parameters for comparison were set to epoch=10, batchsize=16, and imagesize=640. After comparison, the best-performing attention mechanism network structure was selected and adopted. The TJ-Person dataset, as mentioned earlier in this article, was used in this section for the experiments.

As presented in Figure 13, this study compares the effects of introducing different attention mechanisms to the YOLO v5 deep learning model. Five attention mechanisms, namely CA, CBAM, SE, SA, and SK, are compared to demonstrate the effectiveness of the YOLO STP method proposed in this study. Figure 13 (a) shows that the introduction of the SK attention mechanism effectively improves the performance of the model on the train loss and object loss indicators, and the proposed method performs the best among these indicators. Figure 13 (b) indicates that the introduction of the SK attention mechanism improves the F1 Score index of the model in community pedestrian recognition compared to other attention mechanisms during early training. Similarly, the model performance proposed in this study is the best. Figure 13 (c) shows that the introduction of the SK attention mechanism improves the mAP50 and mAP50:90 indicators during early training compared to other attention mechanisms. Although the mAP50 index of the model in this study is slightly lower than that of the SK attention mechanism during early training, it reaches its optimal level after later training. For the mAP50:90 index, the model proposed in this study performs the best. Figure 13 (d) reveals that although the model proposed in this study has an average accuracy in the early stage compared to other models that introduce attention mechanisms, it still achieves highly competitive performance after 10 rounds of training.

In summary, it is evident that the introduction of the SKAttention attention mechanism has a positive effect on the model testing in this study, as indicated by the evaluation of multiple indicators. Moreover, the YOLO-STP model proposed in this paper has achieved the best performance in the evaluation of multiple indicators.



(a)

(b)

(c)

(d)

Figure 13. Comparison of Model Improvement Methods for Multiple Attention Mechanisms, (a) Comparison of Precision Loss; (b) F1 Score curve comparison chart; (c) MAP curve comparison chart; (d) Comparison of detection accuracy curves.

3.3 Comparative testing of pedestrian recognition testing effectiveness in various community scenarios

In order to demonstrate the effectiveness of various attention mechanisms in improving the YOLO v5 model for pedestrian recognition, this study compared the performance of six different methods for detecting pedestrian targets on both the TJ-Person dataset and the COCO-2017 dataset (Kim et al., 2019). A test set of 15 untrained and unverified images was selected from each of the four scenes in the TJ-Person dataset, while in COCO-2017, 20 images were chosen from each of the four scenarios with complex backgrounds, clean backgrounds, light changes, and numerous interference from detection targets. Four representative images were selected for comparison and display. The comparison results of different models are summarized in Table 1, while the detailed comparison results of the test set are shown in Figure 14.

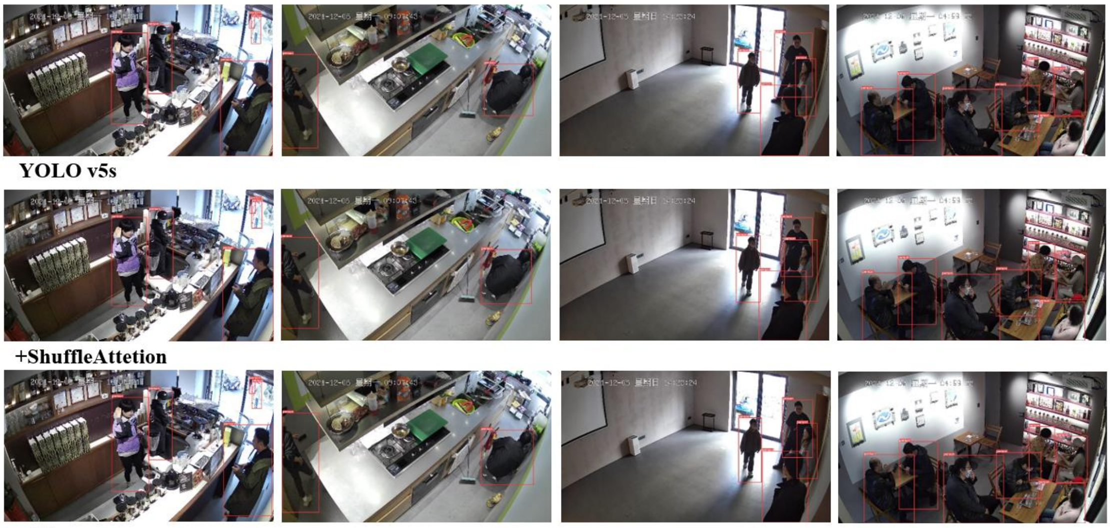
Table 1. Comparison of pedestrian recognition accuracy of models in multiple scenarios using the TJ-Person dataset

Training configuration	**Model improvement methods**	**mAP50(%)**	**mAP50:90(%)**	**F1-Score (%)**	**Optimal detection accuracy (%)**
Dataset：TJ-Person	YOLO v5s	99.08	**80.97**	97.59	97.39
	+ShuffleAttetion	99.03	78.77	97.37	98.05
	+CA	99.03	78.76	97.66	97.89
	+CBAM	99.08	77.94	97.64	97.21
	+SEAttention	98.92	78.73	97.72	97.76
	+SKAttention	99.06	78.74	97.71	97.79
	YOLO-STP	**99.13**	79.00	**97.86**	**98.47**
Dataset：COCO-2017	YOLO v5s	69.28	39.39	67.05	73.82
	+ShuffleAttetion	66.84	36.82	65.10	73.00
	+CA	67.13	37.41	65.56	73.57
	+CBAM	66.81	36.56	65.04	70.66
	+SEAttention	66.46	36.34	64.99	72.64
	+SKAttention	67.98	38.19	66.22	73.18
	YOLO-STP	**70.02**	**40.09**	**67.90**	**74.92**

Based on the experiments conducted on the TJ-Person dataset, it was observed that the SEAttention method had the lowest mAP50 index of 98.92%, while the proposed YOLO-STP method had the highest mAP50 index of 99.13%, which was 0.05% higher than the second-ranked YOLO v5s method. Furthermore, it was found that the CBAM method had the lowest mAP50:90 index of 77.94%, while the YOLO v5s method had the highest mAP50 index of 80.97%, followed by the proposed YOLO STP method with a mAP50 index of 79.00%. The detailed comparison results on the TJ-Person dataset are shown in Table 1, and the comparison of the detection results on the test set is shown in Figure 14.

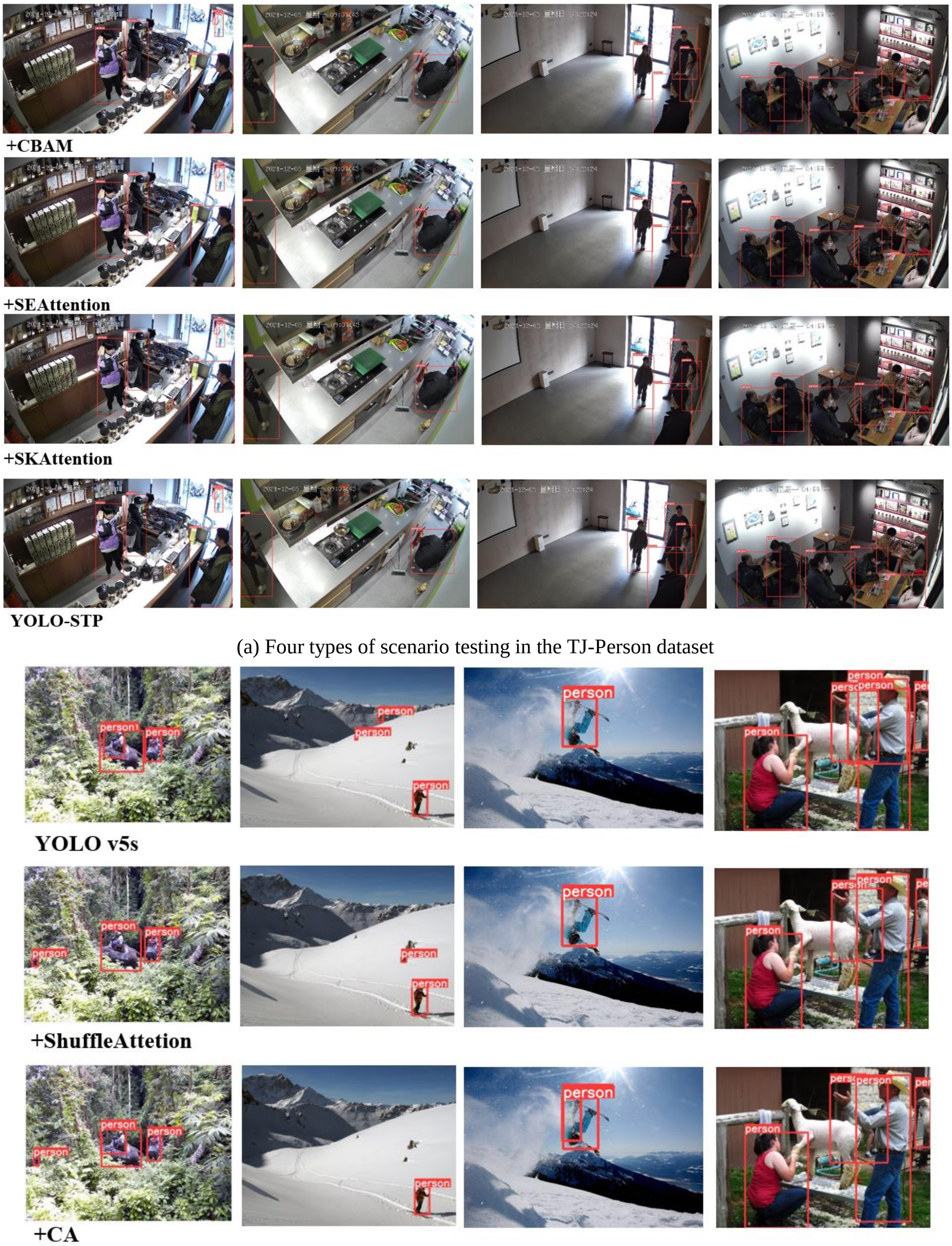
Based on the experiments conducted on both TJ-Person and COCO-2017 datasets, the results indicate that the proposed YOLO-STP method with SKAttention mechanism outperforms other existing methods. The mAP50 index for TJ-Person dataset testing is 99.13%, which is the highest among all tested methods, while the mAP50:90 index for COCO-2017 dataset testing is 40.09%, also the highest among all tested methods. In addition, the YOLO-STP method shows a consistent improvement in performance across all tested indicators compared to other methods. Therefore, the proposed YOLO-STP method can be effectively applied to pedestrian target recognition in open communities.

Overall, the results in Table 1 demonstrate that the proposed YOLO-STP model performs well on both the TJ-Person dataset and the public COCO-2017 dataset, outperforming other improved YOLO v5 models. This indicates that the YOLO-STP model is the optimal choice for pedestrian target monitoring in open communities using the YOLO v5 series.

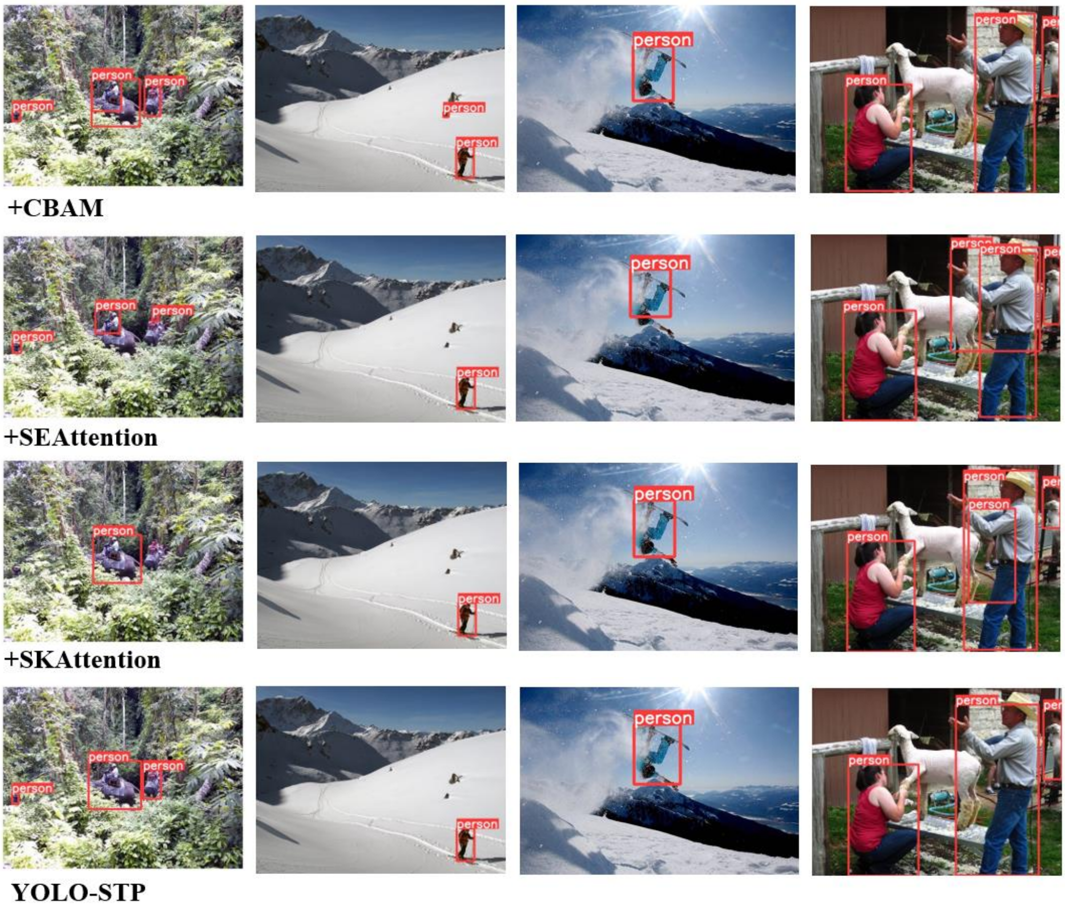

YOLO v5s

+ShuffleAttetion

+CA



(a) Four types of scenario testing in the TJ-Person dataset



(b) Four types of scenario testing in the COCO-2017 dataset

Figure 14. Comparison of YOLO-STP network and various improved YOLO networks in different scenarios. (a) represents the situation of the TJ-Person dataset, and these four types of scenarios are bar, kitchen, activity room, and dining area scenarios from left to right; (b) Representing the situation of the COCO-2017 dataset, these four types of scenes are sequentially classified from left to right as scenes with complex backgrounds, clean backgrounds, varying lighting, and high target interference.

As depicted in Figure 14, our proposed YOLO-STP model consistently achieved excellent results compared to other methods for improving the model when tested on the TJ-Person dataset. This demonstrates that the YOLO-STP model has high reliability in recognizing pedestrian targets in open communities.

Based on the testing results on the COCO-2017 dataset, we can conclude that the YOLO-STP model proposed in this article has superior performance compared to other attention mechanism methods in recognizing pedestrian targets in complex backgrounds and detecting a large number of targets. Meanwhile, it also shows good performance in clean backgrounds and changes in light. On the other hand, the SEAttention and SKAttention methods have flaws in target duplicate detection, and the CBAM method has the lowest mAP50:90 index. These results further demonstrate the effectiveness and reliability of the YOLO-STP model in open community pedestrian target recognition tasks.

In summary, our experimental results under different scenario conditions demonstrate the high reliability and availability of the YOLO-STP method proposed in this article, both in the TJ-Person dataset testing and in the COCO-2017 public dataset testing. It exhibits superior performance in multiple scenarios, indicating the effectiveness and potential of our proposed method in the open community pedestrian target monitoring tasks.

4. CONCLUSIONS

In this article, we conducted a detailed evaluation of the performance of the YOLO-STP network in dealing with pedestrian target recognition problems in open community multi scene spaces. To validate the effectiveness of the method, we constructed a pedestrian target recognition image dataset called TJ-Person, which covers four typical scenarios of a bar, kitchen, activity room, and dining area in the NICE2035 open community in Shanghai. The experimental results show that the improved YOLO STP network achieved an optimal detection accuracy of up to 98.47% in these scenarios, which has significant advantages compared to the original YOLO v5s network and other similar networks.

In addition, we also tested the YOLO-STP network on the COCO-2017 open source dataset to further validate its performance. The experimental results show that compared with other similar networks, the YOLO-STP network has higher recognition accuracy on the COCO-2017 dataset. This result further proves the effectiveness and applicability of YOLO-STP network in pedestrian target recognition in open community multi scene spaces.

However, although the YOLO-STP network has shown high recognition accuracy in experiments, there are still some limitations and potential room for improvement. For example, in scenes with severe occlusion, complex lighting conditions, or high background noise, the performance of the network may be affected. In order to address these challenges, future research can explore more effective attention mechanisms and other advanced computer vision technologies to further improve the performance of YOLO-STP networks in handling pedestrian target recognition problems in open community multi scene spaces.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ACKNOWLEGEMENTS

This work was supported by the Major Program of the National Natural Science Foundation of China (Grant No.42130106) and the “Science and Technology Innovation Action Plan” project of the Science and Technology Commission of Shanghai Municipality (Grant No. 20XD1403800)

The data used in the paper can be obtained by contacting the corresponding author or the following link (https://www.wjx.cn/vm/PQfR5gm.aspx#). The images containing pedestrians used in this article are all from the TJ-Person dataset and the COCO-2017 open source dataset. The TJ-Person dataset has been obtained permission from the photographers, and the COCO-2017 open source dataset is used in accordance with the specifications of the open source protocol.

REFERENCES

Carmona, M., Tiesdell, S., Heath, T., & Oc, T., 2010: *Public Place Urban Space, The Dimension of Urban Design*. Architectural Press/Elsevier, Boston.

Dollár, P., Wojek, C., Schiele, B., Perona, P., 2012. Pedestrian detection: An evaluation of the state of the art. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 34(4), 743-761.

Zhao, H., Shibasaki, R., 2005. A novel system for tracking pedestrians using multiple single-row laser-range scanners. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, 35(2), 283-291.

Zhang, Z., Cui, P., Zhu, W., 2022. Deep Learning on Graphs: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), 249-270.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, A. C., 2016. SSD: Single shot multibox detector. *Computer Vision – ECCV 2016*, 9905, 21-37.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 779-788.

Diraco, G., Leone, A., Siciliano, P., 2015. People occupancy detection and profiling with 3D depth sensors for building energy management. *Energy and Buildings*, 92, 246-266.

Woo, S., Park, J., Lee, J. Y., Kweon, I. S., 2018. CBAM: Convolutional Block Attention Module. *Computer Vision – ECCV 2018*, 11211, 3-19.

Diwan, T., Anirudh, G., Tembhurne, J. v., 2023. Object detection using YOLO: challenges, architectural successors, datasets and applications. *Multimedia Tools and Applications*, 82(6), 9243-9275.

Zhang, Q. L., Yang, Y. bin., 2021. SA-Net: Shuffle attention for deep convolutional neural networks. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, 2235-2239.

Hou, Q., Zhou, D., Feng, J. 2021. Coordinate attention for efficient mobile network design. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 13708-13717.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-Excitation Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, 7132-7141.

Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. Selective Kernel Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 510-519.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B., 2022. Swin Transformer V2: Scaling Up Capacity and Resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 11999-12009.

Liu, W., Liu, Z., Yu, Z., Dai, B., Lin, R., Wang, Y., Rehg, J. M., Song, L., 2018. Decoupled Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, 2771-2779.

Lou, Y., Ma, J., 2018. Growing a community-supported ecosystem of future living: The case of NICE2035 living line. *Lecture Notes in Computer Science*, 10912, 320-333.

Wang, J., Lou, Y., 2022. Exploring the Role of Design in China’s Rural Revitalization Project—The Nice2035 Future Countryside Living Prototyping. *HCII 2022: Cross-Cultural Design. Product and Service Design, Mobility and Automotive Design, Cities, Urban Areas, and Intelligent Environments Design*, 13314, 417-434.

Wang, J., 2021. Social Innovation and Design — Prototyping in the NICE2035 Future Living Labs. *HCII 2021: HCI International 2021 - Late Breaking Posters*, 1498, 71-80.

Kim, D.-H., 2019. Evaluation of COCO Validation 2017 Dataset with YOLOv3. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, 6(7), 10356-10360.