sensors

MDPI

Article

An Improved New YOLOv7 Algorithm for Detecting Building Air Conditioner External Units from Street View Images

Zhongmin Tian [1], Fei Yang [2,3,*] and Donghong Qin [1]

[1] College of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China
[2] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research of Chinese Academy of Sciences, Beijing 100101, China
[3] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
* Correspondence: yangfei@igsnrr.ac.cn; Tel.: +86-10-64889079

Abstract: Street view images are emerging as new street-level sources of urban environmental information. Accurate detection and quantification of urban air conditioners is crucial for evaluating the resilience of urban residential areas to heat wave disasters and formulating effective disaster prevention policies. Utilizing street view image data to predict the spatial coverage of urban air conditioners offers a simple and effective solution. However, detecting and accurately counting air conditioners in complex street-view environments remains challenging. This study introduced 3D parameter-free attention and coordinate attention modules into the target detection process to enhance the extraction of detailed features of air conditioner external units. It also integrated a small target detection layer to address the challenge of detecting small target objects that are easily missed. As a result, an improved algorithm named SC4-YOLOv7 was developed for detecting and recognizing air conditioner external units in street view images. To validate this new algorithm, we extracted air conditioner external units from street view images of residential buildings in Guilin City, Guangxi Zhuang Autonomous Region, China. The results of the study demonstrated that SC4-YOLOv7 significantly improved the average accuracy of recognizing air conditioner external units in street view images from 87.93% to 91.21% compared to the original YOLOv7 method while maintaining a high speed of image recognition detection. The algorithm has the potential to be extended to various applications requiring small target detection, enabling reliable detection and recognition in real street environments.

Keywords: YOLO; heat wave hazards; street view; air conditioner; detection algorithm

check for updates

Citation: Tian, Z.; Yang, F.; Qin, D. An Improved New YOLOv7 Algorithm for Detecting Building Air Conditioner External Units from Street View Images. *Sensors* **2023**, *23*, 9118. https://doi.org/10.3390/s23229118

Academic Editor: Stefania Perri

Received: 9 October 2023
Revised: 1 November 2023
Accepted: 8 November 2023
Published: 11 November 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The frequency and intensity of heat wave disasters are increasing in the context of global warming, posing a severe threat to human health and socio-economic security [1–3]. Therefore, disaster prevention has become critical worldwide. Research has shown that air conditioning is an effective measure for coping with heat wave disasters [4,5]. The 2021 report of the Lancet Countdown on Health and Climate Change revealed that access to effective cooling through air conditioning had already saved tens of thousands of lives in 2019 [6,7].

Understanding the spatial distribution of air conditioner usage is crucial for accurately assessing residents' capacity to cope with heat wave disasters and supporting local governments and policymakers in developing effective policies for disaster prevention and relief. However, traditional statistical methods for collecting air conditioning usage data, whether through direct surveys of residential areas or projections based on local air-conditioning statistics, are costly in terms of human resources, the accuracy of information collection is challenging to ensure, and obtaining spatial distribution data for air conditioning usage in the region is difficult. Thus, there is a critical need to develop a convenient, efficient,

and simple method for extracting information on air conditioner usage and its spatial distribution.

In recent years, street view services like Google Street View, Baidu Panorama, and Tencent Street View have undergone gradual development and improvement. Consequently, the amount of street view image data has significantly increased, and the coverage has expanded to encompass more than 50% of the global population [8,9]. Street view images offer wide coverage and street-level landscape information, serving as a substantial data source and a catalyst for new research ideas in urban environmental evaluation studies. Street view images provide a real depiction of urban streets from a parallel human perspective, which is not attainable through remote sensing data sources like satellite and aerial images [10]. Hence, combining street view images with remote sensing images to extract indicator information represents an ideal approach [11].

Deep learning methods are rapidly advancing in the field of artificial intelligence, providing a means to extract classification features from data sets and achieve automatic and efficient target object classification [12]. Accurate recognition of targets in images is increasingly crucial [13]. In deep-learning-based target detection algorithms, mainstream approaches can be classified as one-stage detection algorithms (e.g., YOLO series) and two-stage detection algorithms (e.g., RCNN series), based on the number of stages in their algorithm structure. The two-stage detection algorithm consists of two target detection processes: candidate region generation and classification. While this approach improves accuracy, it also increases model complexity and time overhead, imposing limitations on computational efficiency [14]. On the other hand, single-stage algorithms directly apply algorithms to input images to generate both candidate regions and target classes through joint decoding. The YOLO algorithm [15], a classic and the first single-stage target detection algorithm, achieves satisfactory results in both detection accuracy and speed by adding classifiers and designing a new loss function to ensure the complete detection of multiple types of targets. Additionally, the YOLO algorithm incorporates advanced feature extraction network designs for efficient detection and lightweight overhead. Being the first deep-learning-based target detection algorithm to support real-time operation, the YOLO algorithm has garnered considerable attention.

The rapid advancements in deep learning methods have spurred a growing number of scholars to explore urban applications utilizing street view image data. Notably in urban planning, the application of deep learning methods to identify the high risk of soft-story buildings in street view images within seismic risk areas has significantly alleviated the challenges associated with traditional manual identification and statistical approaches. This approach enhances work efficiency and provides statistical results to support city managers and decision-makers in formulating strategies to mitigate building risks [16–18]. The fusion of multi-source and multi-modal data, including street view image data, is a widely employed strategy to compensate for the limitations of other data sources and leverage the unique advantages of street view images. This approach significantly enhances accuracy in applied research, particularly in land-use-type classification and building extraction [19–21]. In addition to these macro-level urban planning applications, utilizing street view image data to discern object details is a common strategy for responding to specific urban scenarios. For example, Gebru et al. [22] utilized Google Street View image data to identify the make, model, and year of motor vehicles encountered in a specific neighborhood, totaling 22 million cars (8% of all cars in the U.S.). This analysis accurately estimated income, race, education, and voting patterns at the postcode and precinct level. Krylov et al. [23] utilized Google Street View images to accurately detect the locations of utility poles and traffic lights, achieving a high success rate of over 90% with an accuracy range of 2 m. Chen et al. [24] utilized street view image data from Baidu Panorama to analyze green landscape statistics in various cities in South China. Their research revealed a positive correlation between the number of green landscapes and GDP, as well as other socioeconomic indicators. This highlights the importance of developing public green spaces as the city's economy expands. Nguyen et al. [25] utilized Google Street View image data

to extract diverse indicators related to street greenery, crosswalks, and building types. Through statistical analysis and in conjunction with local health records, they revealed that regions with abundant street and sidewalk greenery exhibited a lower prevalence of obesity and diabetes. Additionally, Hu et al. [26] investigated multiple pedestrian collisions and gathered various road condition variables, including the number of lanes and pavement condition, using street view image data. Their objective was to explore the correlation between collisions and road infrastructure characteristics.

The use of air conditioning, as one of the most important and effective measures to cope with high temperatures, is limited in its role in heat wave disaster response strategies by the difficulty of accurately expressing spatial distribution. Despite the valuable information support provided by street view images from web maps, research on extracting residential air conditioner external units and managing heat wave disasters based on these images is lacking [27]. There are the following difficulties in identifying targets of air conditioner external units in street view image data:

(1) The information extraction algorithms for air conditioner external units have low accuracy;
(2) The street view images exhibit uneven lighting conditions and complex backgrounds;
(3) The varying sizes of air conditioner external units are displayed in the street view image, hindering the observation of small targets from a distance;
(4) The targets are often obscured by objects such as trees and burglar windows, resulting in varying sizes of obscured areas, which severely hinder detection and recognition.

This study aims to propose targeted improvements to the YOLOv7 algorithm [28], which is part of the YOLO series algorithms. These improvements involve the introduction of attention mechanisms in the feature extraction layer and the addition of a small target detection layer [29,30]. This aims to solve the difficult problems of severe occlusion, complex background, low accuracy, and difficulty in identifying small targets in the application of identification of air conditioner external units. By developing an improved YOLOv7 algorithm, this study aims to enhance the accuracy of detecting air conditioner external units in vast street view images. These improvements aim to provide methodological and data support for rapidly and accurately acquiring the spatial distribution of air conditioner use. They also contribute to the development of more rational heat wave disaster response tactics and ultimately enhance the efficiency of such responses.

2. Materials and Methods

2.1. Study Area and Data

Currently, Baidu Maps has collected panoramic data for streets in 652 cities in China, covering a mileage of 2,295,000 km, and in most cities, users can access the street view of the roads in Baidu Maps. In the developer mode of Baidu Maps, developers can use the API interface after converting the coordinate system according to the coordinate point to obtain the target location of the 360° street panoramic image or other arbitrary direction of any angle range of the street image—usually, the distance from the setting of the road network coordinates every 25 m positioning, and thus developers can obtain a complete, coherent street image data.

The reproduction of a 360° panoramic image results in a degree of spatial distortion, which causes a distorted panoramic street view image (Figure 1a); as there is the choice of every 90° angle of four directions of the frontal street view image data, the image data obtained are more closely related to the real situation of the street scene (such as Figure 1b–d). Since the purpose of this study is to detect and count the air conditioner external units hanging on the walls of buildings, it is sufficient to choose only two directions of 90° and 270°, i.e., the left and right directions of the road.



(**a**) Whole 360° panoramic street view image.



(**b**) 90°direction



(**c**) 180°direction



(**d**) 270°direction



(**e**) 360°direction

Figure 1. Example of frontal street view image orientation map.

This study focuses on the urban area of Guilin City, located in the Guangxi Zhuang Autonomous Region of China, as the study area for extracting air conditioner external units from street view image data on buildings along the roads. Guilin City, a renowned tourist destination, is situated in a subtropical region and experiences recurring high-temperature heat waves annually, posing a substantial threat to both local population and international visitors. As one of the initial cities chosen for China's Sustainable Development Demonstration, Guilin is in immediate need of conducting research on high-impact urban disaster response and sustainable development. The construction of target detection models heavily relies on rich and rigorous data sets [31]. The street view images predominantly cover the main roads in Guilin City, with fixed points established every 25 m in the continuous urban main road network data, resulting in a total of 61,929 distinct road network coordinate points. By utilizing the acquired coordinate point information, a bulk collection of 123,858 street view images from both sides of the road was achieved through the Baidu panorama service. Subsequently, redundant street view images displaying excessive similarity were manually eliminated, resulting in the selection of 5680 instances of air conditioner external units as the training data set. The data set is comprehensive in its coverage, containing individual examples of air conditioner external units in a variety of states, including those obscured by trees and windows, those with small targets at a distance, and those placed in an oblique direction. In order to control the comprehensiveness and balance of the sample data set, we also manually screened the samples so that the number of samples of air-conditioning outboard units in each different state could occupy a large proportion. The screened street view images are annotated using the graphical image annotation tool 'labelImg' (https://github.com/tzutalin/labelImg, accessed on 8 October 2023). This tool identifies and frames all the air conditioner external units in each image, generating an XML file that contains the target type and coordinate information. The annotated data are then split in an 8:1:1 ratio for algorithm training, testing, and validation, respectively.

2.2. Data Pre-Processing and Data Augmentation

The 25 m distance interval set for the coordinates of the urban trunk road network ensures the continuity and integrity of the street view image data. However, in open areas, where there are few or no buildings nearby and highly overlapping scenes in the distance, it is necessary to pre-process the data by removing street view images that exhibit excessive similarity. This is completed to ensure data accuracy, minimize redundancy, and improve data utilization efficiency. Future research utilizing Street View image data will enable more accurate analysis and understanding of urban development.

Upon completing the data pre-processing, data augmentation is performed on the filtered data to mitigate potential overfitting issues. Data augmentation expands limited data to create more diverse training samples, thereby increasing both sample size and diversity, and improving model robustness. Given the characteristics of street view images, which may feature objects at varying distances, complex backgrounds, diverse lighting conditions, and occlusions, the Mosaic data augmentation method proposed in YOLOv4 [32] was chosen for this study. This approach combines four images into a single training sample, effectively reducing GPU memory usage and enhancing network robustness without requiring large mini-batch sizes. By randomly scaling and distributing images for stitching, Mosaic data augmentation enabled the synthesis of new image data with numerous small targets, and edge location labels assisted in identifying occluded objects, thereby enriching the detection data set. Figure 2 illustrates the results of this method.



Figure 2. Mosaic data augmentation effect ("0" is the classification number of the air conditioner external unit).

2.3. Performance Evaluation Index

The performance evaluation indices consist primarily of mean average precision (MAP), precision (P), and recall (R). These metrics are calculated via the following formulas:

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{1}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{2}$$

$$AP = \int_0^1 P(R)\, dR \tag{3}$$

where, T_P represents the actual positive samples that were detected as positive, F_P represents the actual negative samples that were erroneously detected as positive, and F_N represents the actual negative samples that were correctly detected as negative. In turn, P corresponds to the ratio of correctly predicted outcomes to all predicted positive cases, whereas R pertains to the ratio of correctly predicted outcomes to all positive cases. The area under the P-R curve represents the AP value, with the mAP value calculated as the average AP values across all categories. To evaluate the detection performance of the complete target detection network model, the mAP value is commonly employed. For mAP@0.5, IoU (Intersection over Union) is established at 0.5, with the AP of all images in every category calculated and averaged across all categories. Similarly, mAP@0.5:0.95 is

representative of the average mAP across distinct IoU thresholds (ranging from 0.5 to 0.95 in increments of 0.05).

The NMS method may not yield entirely accurate prediction frames and corresponding categories. Therefore, samples with confidence levels exceeding threshold a = 0.5 are identified as positive, and vice versa. Moreover, samples that overlap actual frames by an intersection ratio surpassing threshold d = 0.6 are designated as true positives (T_P) while those that do not comprise false positives (F_P). If any actual positive samples exist within the negative samples, they are categorized as false negatives (F_N).

2.4. YOLO

This study selected three versions of the "You Only Look Once" (YOLO) series, namely YOLOv5, YOLOv7, and YOLOv8, as the primary algorithm choices. These single-stage target detection algorithms are preferred in current engineering applications due to the relative inefficiency of two-stage target detection algorithms like RCNN [33,34]. The experiments were conducted using a pre-processed data set of air conditioner external units. The detection results are presented in Table 1. It is evident from the results that YOLOv5 exhibits a slightly lower detection performance than the two latest versions, while the superiority of YOLOv8 over YOLOv7 is not significant. However, due to its recent release, YOLOv8 still lacks stability, and there are more unknown errors associated with implementing improvements and modifications compared to YOLOv7.

Table 1. Comparison of evaluation indexes of detection effect of basic algorithms.

Algorithm	P	R	mAP@0.5	mAP@0.5:0.95
YOLOv5	0.8760	0.7389	0.8441	0.5080
YOLOv7	0.9024	0.8397	0.8793	0.5466
YOLOv8	0.9053	0.8434	0.8829	0.5487

Therefore, this study utilizes YOLOv7 as its foundation, which is the most reliable iteration of the YOLO series to date. The ELAN network [35], an efficient aggregation network, is employed within the YOLOv7 architecture to optimize memory access costs and GPU computing efficiency. In order to prevent information over-inflation and loss during multi-layer propagation, the ELAN network facilitates interaction between layers, enabling deeper network depth and improved accuracy. The ELAN network structure controls connection paths of varying lengths to promote effective network learning and convergence. YOLOv7 incorporates module-level reparameterization techniques that break modules into several equivalent microstructures with consistent parameter transformations, which greatly enriches training resources and enhances model performance while maintaining prediction consistency with previous YOLO iterations. The YOLOv7 architecture retains the three-part structure of its predecessors—including the backbone, neck, and YOLOHead—and is illustrated in Figure 3.

The backbone network serves as the primary feature extractor in YOLOv7, extracting image inputs and generating corresponding feature layers. Three effective feature layers are retained in the backbone section for use in subsequent network construction. These layers represent the feature sets of input images and serve as a critical element to inform network performance.

The neck network, which acts as an enhanced feature extractor, merges the three effective feature layers generated in the backbone section in order to amalgamate feature information from various scales. The Panet structure remains employed in YOLOv7, enabling both up-sampled and down-sampled feature fusion. In the neck section, feature extraction continues through utilization of the effective layers obtained previously.

Finally, the YOLOHead function serves as both classifier and regressor, enabling access to the three enhanced effective feature layers generated by the backbone and neck. Specifically, YOLOHead assesses whether a priori boxes correspond to objects detected in

feature points. As with earlier iterations of YOLO, the decoupling head remains integrated in YOLOv7, executing classification and regression through a 1×1 convolution approach.

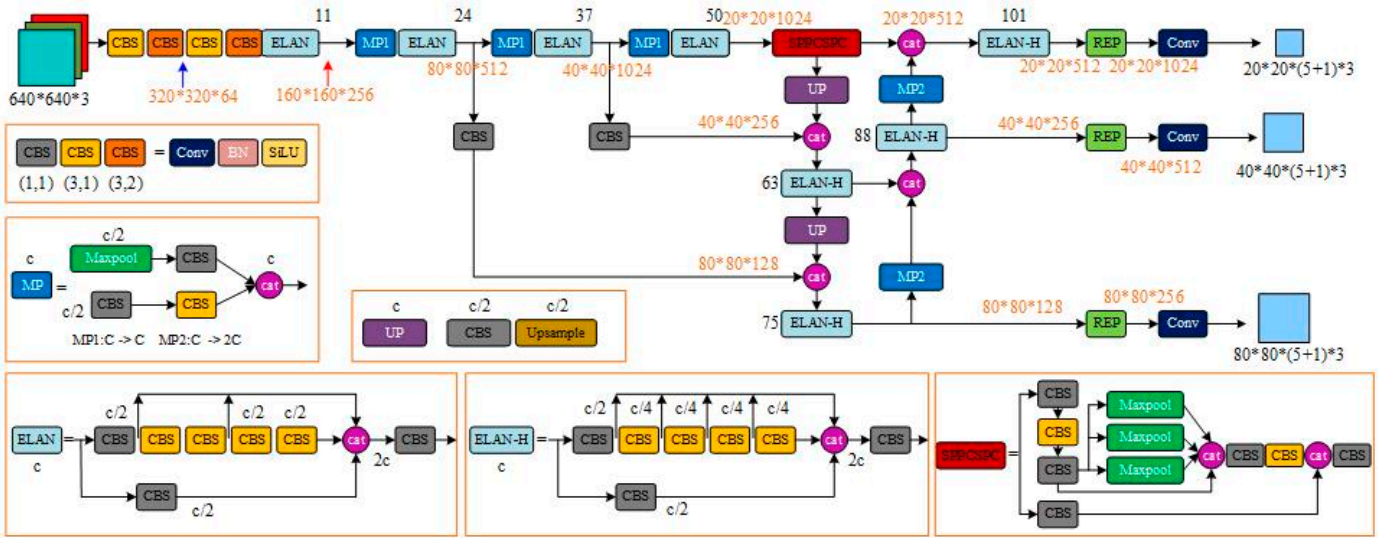


Figure 3. Original YOLOv7 network structure diagram.

Despite the performance improvements in YOLOv7 compared to its predecessors, it still faces challenges in detecting small targets and performing target detection tasks in complex backgrounds. Consequently, it fails to meet the specific requirements of this study, which focuses on detecting air conditioner external units in street view image data. Therefore, targeted improvements are necessary.

3. Results

3.1. SC4-YOLOv7 Algorithm Improvement

The YOLOv7 algorithm is a new algorithm of the YOLO series proposed in 2022 and is a one-stage detector with very good performance, so it is chosen as our base model. YOLOv7 requires pre-training weights on large-scale data sets and is less effective in recognizing air conditioner external units in complex backgrounds in street view image data. Therefore, in this study, we optimize and improve the initial YOLOv7 model in three aspects, namely, the backbone network, enhanced feature extraction network, and detection head, to form a new algorithm for efficient extraction of air conditioner external units in complex background environment of street view image data, in response to the small size of air conditioner external unit targets in street view images and the fact that they are easily obscured and other features are not obvious.

3.1.1. Backbone Network Optimization—Introducing SimAM Parameter-Free Attention

The visual attention mechanism, a unique signal processing mechanism in the human brain, is an inherent component of human vision. It enables rapid scanning of the global image to identify the area requiring focus, commonly referred to as the attention focus. Subsequently, it allocates additional attention resources to this area, enabling the capture of more detailed information while suppressing peripheral, irrelevant details. Given the complex characteristics of street view image data, this algorithm incorporates an attention mechanism module to mitigate the impact of complex backgrounds on detection results. This approach helps the model prioritize the extraction of air conditioner external unit features and enhances learning regarding these features within relevant image regions.

While some existing attention modules refine features solely in either the spatial or channel dimensions, their limited flexibility becomes evident when both undergo simultaneous change. The SimAM module addresses this challenge by serving as a parameter-free 3D attention module that does not require additional parameters to be added to the original

network. In contrast to one-dimensional (1D) channel attention and two-dimensional (2D) spatial attention modules, which focus only on channel or spatial location importance, respectively, 3D attention enables simultaneous channel and spatial location feature attention. This approach infers 3D attention weights by analyzing the feature mapping, allowing for consideration of both channel and spatial location importance simultaneously.

SimAM 3D weights implement a visual neurological theory that acknowledges neurons' unique firing patterns and spatial depression effects in visual processing tasks. Neurons demonstrating spatial depression effects should be assigned higher weights, and the easiest means of identifying these neurons is measuring linear differentiability between the target neuron and its surrounding ones. Thus, an energy function is defined for each neuron based on said scientific findings. If an input image R with input features $X \in R$ $C \times H \times W$ exists in C channels and $M = H \times W$ neurons, each channel can theoretically have M energy functions. However, to reduce iterative computation and avoid the need to calculate the mean and variance for each position, the average mean μ and variance σ of all neurons x_i from a single channel can be calculated using Equations (4) and (5).

$$\hat{\mu} = \frac{1}{M} \sum_{i=1}^{M} x_i \tag{4}$$

$$\hat{\sigma}^2 = \frac{1}{M} \sum_{i=1}^{M} (x_i - \hat{\mu})^2 \tag{5}$$

Therefore, the final minimum energy function is shown in Equation (6):

$$e_t^* = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{6}$$

where t is the target neuron and λ is the hyperparameter.

From Equation (3), SimAM can be evaluated for each neuron of each network by defining the energy function of linear differentiability and calculating the minimum energy e_t^*. The lower the energy of e_t^*, the more the neuron t is distinguished from the surrounding neurons and the higher the importance. Therefore, the importance of neurons can be obtained by $1/e_t^*$. Finally, based on the gain effect of neuron response, the weighting is performed by using the deflation operation, while the sigmoid function can also limit the excessive values of E (features grouped by neuron importance $1/e_t^*$) but also does not affect the relative importance of each neuron, as in Equation (7).

$$\widetilde{X} = sigmoid\left(\frac{1}{E}\right) \odot X \tag{7}$$

When recognizing image features, both channel and spatial location are essential, and therefore, this paper introduces parameter-free attention SimAM to the ELAN of the backbone network to evaluate feature importance comprehensively and effectively. By leveraging the energy function, features extracted from the backbone can be evaluated, and important neurons carrying rich messages can be identified. This improved model enables the identification of crucial features, the suppression of irrelevant feature interference, improvement in the network's feature representation capability, and enhancement of the model's target localization ability to focus on feature representation that facilitates recognition of air conditioner external unit targets in street view image data.

3.1.2. Neck Network Optimization—Introducing Coordinate Attention

The Coordinate Attention (CA) mechanism efficiently embeds location information into channel attention, allowing the CA module to pay attention to both regions of interest in the channel and their corresponding locations. This two-step approach involves information embedding and attention weight generation.

Traditional attention mechanisms, including SENet [36], ECA [37], CBAM [38], etc., usually use global pooling to globally encode spatial information, which ignores location information. In order to accurately capture both channel information and location information, the information embedding stage first uses pooling kernels of size (H, 1) and (1, W) for each channel in the horizontal and vertical directions, respectively, for the feature layers input to the CA module. Better capture of location information and channel information with a small increase in computational effort facilitates the network to locate the target of interest. The pooling outputs for the cth channel with height H and the cth channel with width W are as follows:

$$Z_C^h(h) = \frac{1}{W}\sum_{0\le i\le W} x_c(h,i) \tag{8}$$

$$Z_C^w(w) = \frac{1}{H}\sum_{0\le j\le H} x_c(j,w) \tag{9}$$

The features in the horizontal and vertical directions of the feature map are represented by $x_c(h,i)$ and $x_c(j,w)$, respectively. The horizontal and vertical tensors, $Z_C^h(h)$ and $Z_C^w(w)$, respectively, are obtained after pooling. These tensors are then stitched together and transformed using a 1×1 convolution function, F_1, as follows:

$$f = \delta\left(F_1\left(\left[Z^h, Z^w\right]\right)\right) \tag{10}$$

In the given equation, δ is the nonlinear activation function, [..., ...] denotes the splicing operation, and f is the intermediate feature for encoding spatial information in the horizontal and vertical directions. The f is decomposed into two separate tensors f_h and f_w in horizontal and vertical directions, and finally the number of channels of the two tensor is made consistent by a 1×1 convolution F. Using the sigmoid activation function σ outputs,

$$g^h = \delta(F(f_h)) \tag{11}$$

$$g^w = \delta(F(f_w)) \tag{12}$$

The output results g^h and g^w of the above equation are the weights of attention in the vertical and vertical directions, respectively. The initial input feature $x(i,j)$ is multiplied with the weights of the corresponding positions to obtain the output result $y(i,j)$, and the final output of the CA module is as follows:

$$y(i,j) = x(i,j) * g^h(i) * g^w(j) \tag{13}$$

The objective of our study is to enhance the YOLOv7 network's ability to capture target features in street view image data of air conditioner external units located outside. To achieve this goal, a Coordinate Attention module is incorporated into the network's neck. The inclusion of this module ensures that the network focuses on the location and channel information of the relevant target features, thus enabling better feature extraction.

3.1.3. Head Network Optimization—Add Small Target Detection Layer Head

Compared with conventional targets, small targets are characterized by weak features and little information, and it is very difficult to distinguish them from similar backgrounds or contiguous targets. When facing the problems of complex environments such as low illumination and shadow occlusion, the task of detecting small targets in street view image data poses a higher challenge.

The original YOLOv7 model backbone network was downsampled a total of five times to obtain five layers of feature expressions (P1, P2, P3, P4, and P5), where Pi denotes a resolution of 1/2i of the original image, and although multi-scale feature fusion was achieved in the neck network through top-down and bottom-up aggregation paths, it did not affect the scale of the feature maps, and the final detection head was partially in

the detection head induced through P3. The detection of the target is performed on the detection heads led through the three-level feature maps P3, P4, and P5, whose feature map scales are 80 × 80, 40 × 40, and 20 × 20, respectively. For the convenience of expression, the detection heads led through the Pi-level feature maps are referred to as Pi-level detection heads hereafter.

The self-built street view image data discussed in this paper contain more distant individual targets of air conditioner external units, whose scales are often smaller than 20 × 20 pixels. These targets lose most of their feature information after repeated downsampling, and thus it remains challenging to detect them even by the higher resolution P3 layer detection head. To address this issue and achieve better detection results for these tiny targets, we introduce a new detection head on the YOLOv7 model that utilizes the P2 layer features. The resolution of the P2 layer detection head is 160 × 160 pixels, equivalent to only two downsampling operations in the backbone network, containing richer underlying feature information of the targets. The two P2 layer features obtained from top-down and bottom-up paths in the neck network are fused with the same scale features in the backbone network through concatenation. The output features are the fusion results of the three input features, making the P2 layer detection head fast and effective in dealing with tiny targets. Moreover, the newly added detection head is specific to the underlying features and generated from a low-level, high-resolution feature map that is more sensitive to tiny targets. Although the addition of this detection head increases the model computation and memory usage, it effectively mitigates the negative impact of scale variance. The P2 detection head, together with the original three detection heads, can effectively improve the detection capability for tiny targets.

Through practical tests, the actual results also prove that after adding the small target detection layer, the model detects a significant increase in the individual detection rate of small targets in street view image data.

3.1.4. Building the New SC4-YOLOv7 Algorithm

In this paper, we propose a new SC4-YOLOv7 model that is optimized and improved in three parts: the original YOLOv7 model backbone network, enhanced feature extraction network, and detection head. The purpose of these changes is to make the model more effective in detecting targets in complex background environments and tiny targets in street view image data. Figure 4 shows the structure of the network after these modifications.

The SC4-YOLOv7 model comprises four parts: input, backbone network, neck network, and detection head. The input part enhances the training data with techniques like Mosaic and applies adaptive scaling and anchor frame calculation to images. To address issues such as similar backgrounds in street view images, difficult feature extraction due to texture replication, and loss of network propagation feature information, we introduced the SimAM attention module to the backbone network of the original YOLOv7 model. This module utilizes an energy function to identify effective features and suppress irrelevant ones, thereby improving the feature extraction capability of the backbone network for small targets. The neck network is based on the SPP + PAN structure but includes a coordinate attention mechanism for feature fusion to obtain hierarchical feature representations that are passed on to the detection head. A small target detection layer is added to the detection head to provide four detection scales (10 × 10, 20 × 20, 40 × 40, and 80 × 80). Finally, the optimal prediction frame is determined by non-maximal suppression (NMS) of the target detection frame.

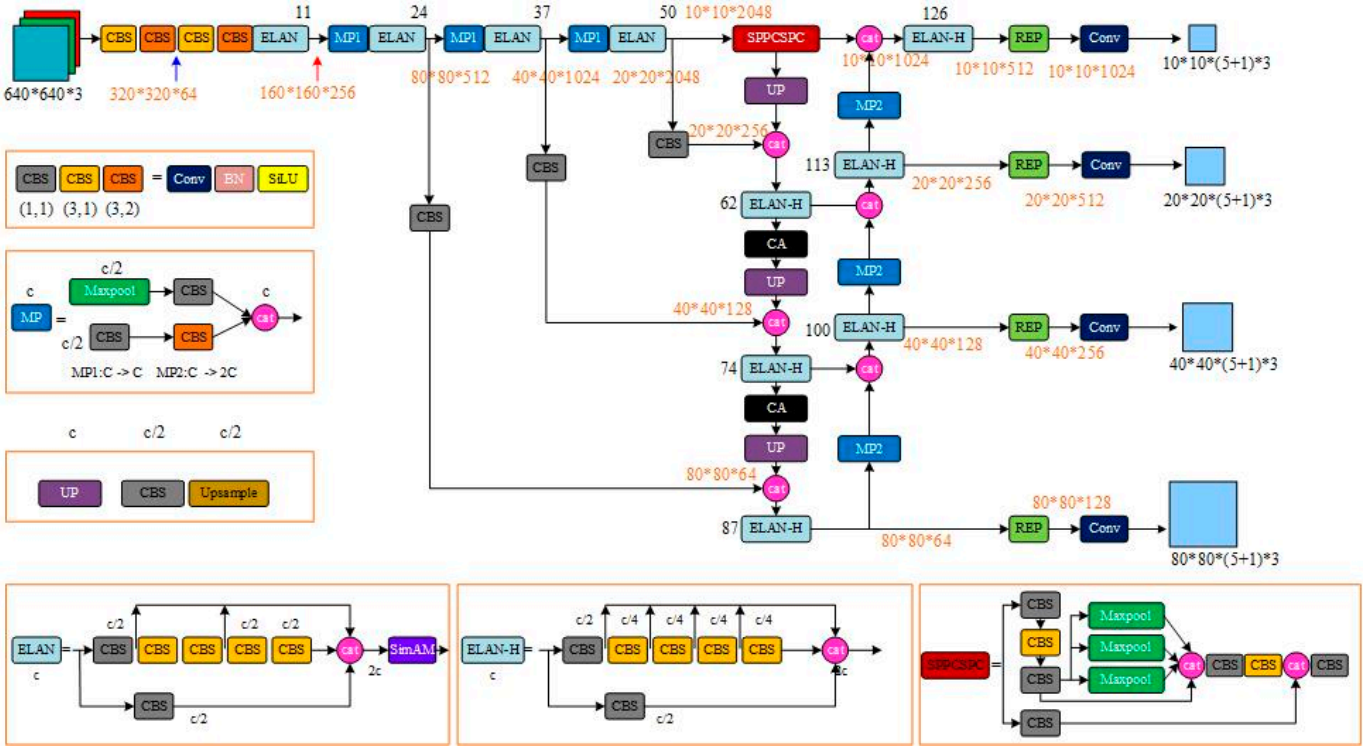


Figure 4. SC4-YOLOv7 network structure diagram.

3.2. Research on Extraction of Air Conditioner External Units from Street View Images

3.2.1. Experimental Environment Configuration

In this study, all comparison models perform detection operations on the GPU server. Table 1 shows the experimental configuration. A total of 5680 images of individual air conditioner external units were used for training during the detection process. The stochastic gradient descent (SGD) momentum of all target detection algorithms in this experiment was set to 0.9. The initial learning rate is set to 0.01, the weight decay is 0.0005, and the training is performed using a pre-training strategy. The input size is fixed to the same size as used in the detection process (Table 2).

Table 2. Experimental configuration.

Configuration	Parameter
System Environment	ubuntu20.04
GPU	V100-SXM2-32GB
Accelerated environment	CUDA 11.3
Libraries	PyTorch1.11.0
Language	Python 3.8

3.2.2. Evaluation of the Performance of Air Conditioner External Unit Extraction of Street View Image Data

The experimental comparison indicates that the small target detection layer improves the recall from 0.8397 to 0.8588, mAP@0.5 from 0.8793 to 0.8945, and mAP@0.5:0.95 from 0.5466 to 0.5681 in the original YOLOv7 model despite a slight reduction in precision. With the addition of SimAM and CA attention mechanisms, the model's comprehensive detection ability improves with little effect on recall rate but with increases in precision, mAP@0.5, and mAP@0.5:0.95 indexes from 0.9024, 0.8793, and 0.5466 to 0.9198, 0.9017, and 0.5683, respectively. By combining the small target detection layer with SimAM and CA

attention mechanisms, the model exhibits further improvement in detection ability, with precision, recall, mAP@0.5, and mAP@0.5:0.95 indexes reaching 0.922, 0.8660, 0.9121, and 0.5977 (Table 3). The model performs well in detecting air conditioner external units, even under challenging conditions such as large target individuals at close distances, small target individuals at long distances, and individuals blocked by trees and fences. Compared with the original YOLOv7 algorithm, the new SC4-YOLOv7 algorithm shows significant improvements in detecting target objects and can detect many objects that were previously missed. Figure 5 provides a comparison of the training process. Figure 6 provides a comparison of the results.

Table 3. Comparison of evaluation indexes of detection effect of each model.

Algorithm	P	R	mAP@0.5	mAP@0.5:0.95
YOLOv7	0.9024	0.8397	0.8793	0.5466
YOLOv7 + SimAM + CA	0.9198	0.8317	0.9017	0.5683
YOLOv7 + 4Head	0.8822	0.8588	0.8945	0.5681
SC4-YOLOv7	0.9220	0.8660	0.9121	0.5977

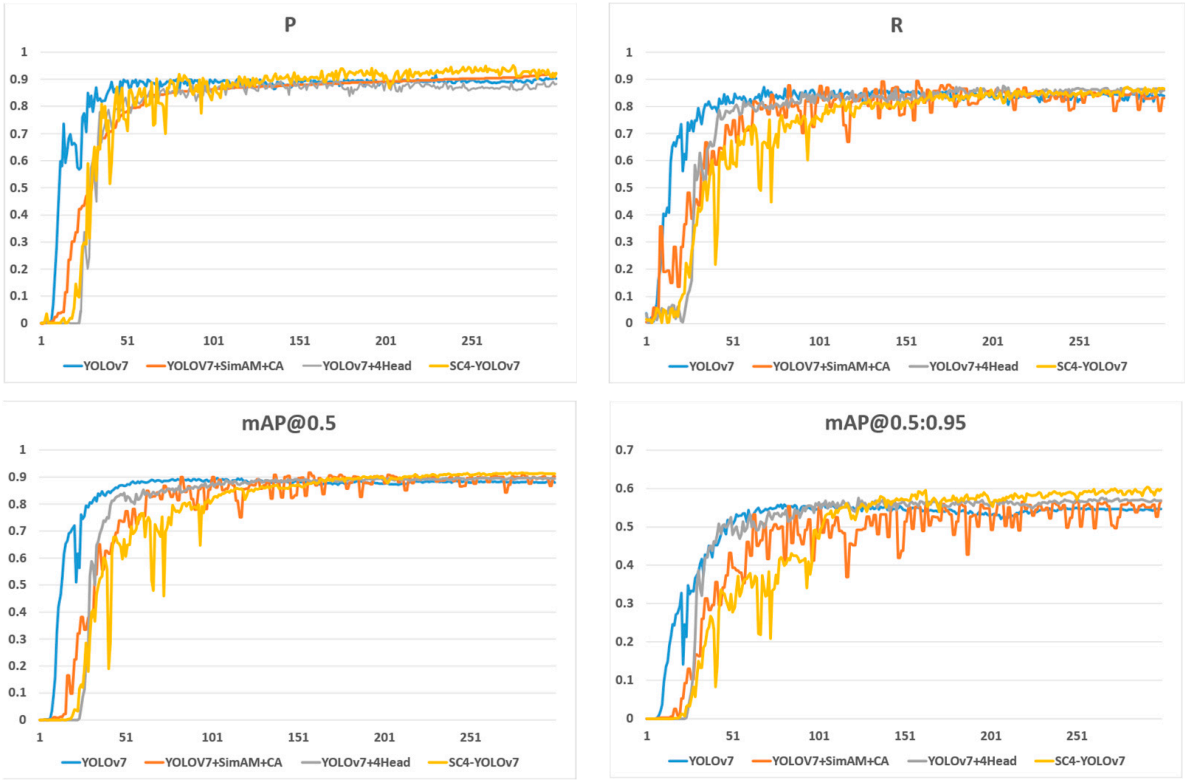
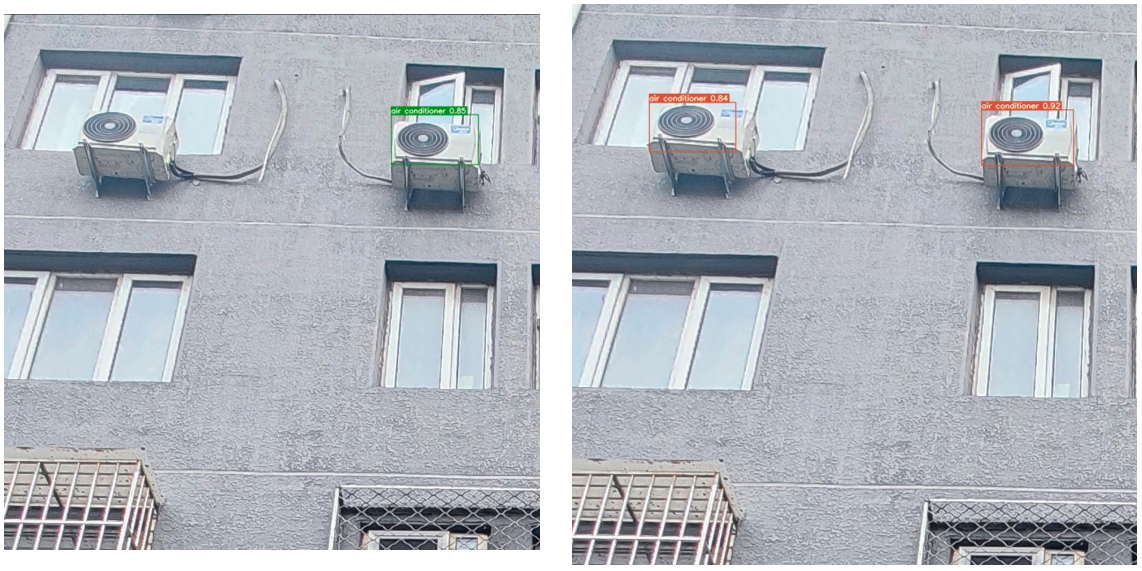


Figure 5. Each algorithm detects the experimental effect.

The SC4-YOLOv7 network model was used to detect the air conditioner external units on all the acquired street view image data of the main roads in Guilin, Guangxi Zhuang Autonomous Region, and the results achieved the expected goal: the air conditioner external units were accurately detected. Some of the results are shown in Figure 7.



(**a**) Original YOLOv7 target detection effect (**b**) SC4-YOLOv7 new algorithm target detection effect

Figure 6. Comparison chart of the effect of the target detection task of air conditioner external unit.



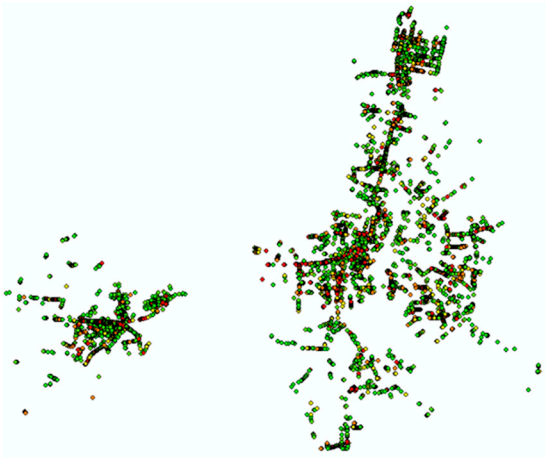
Figure 7. The effect of detecting the air conditioner external unit target under different conditions such as obscured by fence, obscured by tree, oblique direction position, and small target from a distance in street view image data.

3.3. Spatial Distribution of Urban Air Conditioner External Units

The new data combination is formed by counting the number of air conditioner external units detected in each street view image and combining the corresponding coordinate information. Using ArcGIS, an air conditioning statistical map is created based on geographical location and the number of detected frames, which corresponds to the number of detected air conditioner external units. In Guilin city, there are 61,929 coordinate points of road networks and 123,858 street view image data. The use of the YOLOv7 algorithm results in the detection of 13,806 street view image data, 11,752 coordinate points, and

87,176 air conditioner external units, while the YOLOv7 + SimAM + CA algorithm detects 13,873 street view image data, 11,805 coordinate points, and 99,234 air conditioner external units. Furthermore, using the YOLOv7 + 4Head algorithm, 13,959 street view images, 11,869 coordinate points, and 114,118 air conditioner external units are detected, while the SC4-YOLOv7 algorithm detects 13,933 street view images, 11,850 coordinate points, and 113,553 air conditioner external units. Notably, the SC4-YOLOv7 algorithm exhibits superior performance in detecting 113,553 air conditioner external units in 13,933 street view images with 11,850 coordinate points. Based on the detection statistics of each algorithm, a distribution map of air conditioners is generated, and part of the situation is shown in Figure 8.

As can be clearly seen from the detection results point diagram, when the original YOLOv7 algorithm joins the small target detection layer, the network can pay more attention to the tiny target individuals and the checking of all rates by the algorithm increases significantly, so the image indicates a significantly greater number of orange and red dots. However, relatively, the algorithm accuracy will have a certain degree of reduction, that is, a large number of small target individuals that are not air conditioner external units will be used in air conditioner external unit detection. The YOLOv7 network with the attention mechanism can better focus on the required air conditioner external unit features from the complex background environment of the street view images, correctly detect and identify them, and improve the accuracy of the algorithm. The SC4-YOLOv7 algorithm in this study combines the above advantages and is generally optimal considering the accuracy of the detection under the condition of ensuring few or even no omissions.



(**a**) YOLOv7



(**b**) YOLOv7 + SimAM + CA



(**c**) YOLOv7 + 4Head



(**d**) SC4-YOLOv7

Figure 8. *Cont.*

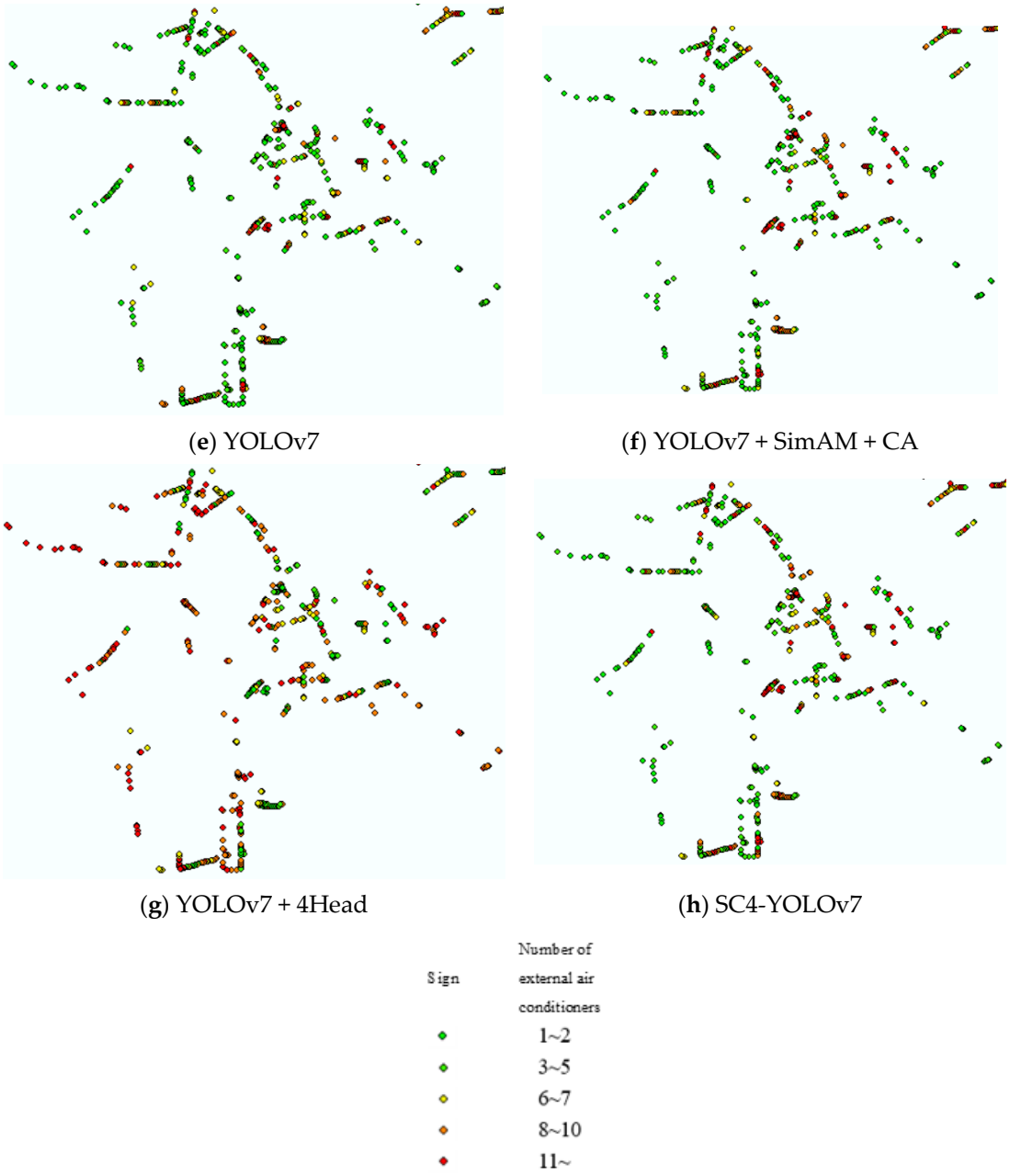


Figure 8. Air conditioner external space distribution map.

The images obtained by combining the coordinate data of the main road network of Guilin city and the statistical data of the number of target detection of air conditioner external units in street view images can reflect the high coverage and concentrated distribution of local air conditioners very intuitively and provide corresponding data support for the policy formulation and method implementation of heat wave disaster prevention.

4. Discussion

Street view image data provide researchers with a low-cost, efficient, and wide-coverage way to record and observe the physical urban environment from a parallel human perspective [39]. In combination with the rapidly developing field of artificial intelligence, the ability to collect, process, compute, and analyze large-scale data in a short period has increased dramatically and has now become an important data source for urban application research, bringing new methods and ideas [40].

The assessment and analysis of air conditioner coverage and utilization, serving as crucial indicators in the context of addressing urban heat wave disasters, primarily rely on outdated manual techniques. Street view image data, which offer a vibrant reflection of a city's status, have been overlooked as a potential source of information. Currently, available artificial intelligence algorithms, catering to a broad spectrum of applications, have not been specifically implemented for the targeted recognition of air conditioner external units in street view image data. Consequently, this study presents a deep-learning-oriented SC4-YOLOv7 algorithm for the precise detection and counting of air conditioner external unit targets on buildings within street view image data. This approach offers a more convenient and efficient solution for related research endeavors.

Although this study only focuses on the target detection and counting statistics of air conditioner external units in urban street view image data, the overall method can be fully migrated to the detection of other target objects in street view image data with complex background environment conditions that are also obscured by trees and fences and small targets of different sizes and distances that are difficult to detect, and the idea of applying street view image data efficiently and combining them with location coordinate information will support future multi-scale and multi-dimensional urban information mining and extraction, and urban residential environment enhancement.

As street view image data are gathered by the street view collection vehicles, equipped with panoramic equipment, while traversing city roads, the current coverage of both Google Street View and Baidu Panorama services remains limited. At this stage, the street view image data primarily cover major city roads. However, data from smaller roads and residential communities, where vehicles face difficulty navigating, are essentially unavailable, which has also led to the inability to complete detail-demanding and relevant urban studies based on Street View image data alone. Our study's street view image data are based solely on the information of the city's main road network in Guilin City, which unfortunately cannot comprehensively represent the entire urban environment. Nonetheless, the counting results from our research provide a sufficient macroscopic perspective to reveal the proportional density of air conditioning coverage and its spatial distribution across the city. Moving forward, we will continue to refine and expand the data scope by covering areas that are currently excluded, such as selecting residential communities and deploying professionals to take additional photographs of air conditioner external units located on interior building walls, among other improvements.

However, street view image data are inherently limited as they can only capture the physical environment that is visible to the human eye in urban areas. These data struggle to effectively represent natural environment characteristics such as urban air, soil, and water conditions. Therefore, future studies focusing on urban applications based on street view images will undoubtedly require the integration of multiple data sources, including remote sensing images, geotagged social media data, and more. Remote sensing images, obtained through remote sensing technology, offer an overhead view of the ground [41–43], enable the visualization of landform shapes, and possess the advantages of providing a macroscopic, timely, and rapid acquisition of features. However, current urban applications that combine street view images and remote sensing data largely focus on urban greening, a domain that is nearing saturation. There is a significant lack of research addressing other natural environments. Geotagged social media data reflect the environmental situation from both rational and emotional perspectives [44–46], and when combined with street view image data, offer a more comprehensive analysis of target areas. However, the dispersed and voluminous nature of social media data with geographical locations presents challenges in data extraction and filtering, limiting the number of studies that incorporate these data.

In our future work, we plan to utilize data sources that are not limited to web-based street view images. Specifically, we aim to integrate multiple data sources to expand our analysis capabilities, such as incorporating subjective social media photos with geographic location data in conjunction with street view images. Additionally, we will leverage data

sources containing more subjective emotions, such as questionnaires, to better understand the direct impact of environmental factors on residents. Remote sensing image data serve as a complementary source to street view image data, offering higher-level and more macroscopic information. We propose combining remote sensing data with natural data sources [47], including geothermal and air environment data, as well as humanistic data sources, such as AOI, rooftop, and pedestrian flow data, to enhance urban multi-source data mining, information extraction, and provide powerful support for urban disaster prevention and mitigation, environmental analysis, rational planning, and intelligent construction.

5. Conclusions

In this study, we utilized continuous street view image data collected from both sides of the city's main roads to detect air conditioner external units on buildings. We identified the limitations of the original YOLOv7 algorithm model in extracting information about these external units and addressed them. To improve the detection accuracy of air conditioner external unit recognition for street view image data under various complex environmental conditions, we developed a new algorithm called SC4-YOLOv7. This algorithm incorporates several enhancements, including the addition of SimAM and CA attention mechanisms, as well as the inclusion of small target detection layers.

Compared to the original YOLOv7 algorithm, the proposed SC4-YOLOv7 algorithm demonstrates significant improvements in detecting air conditioner external unit targets in street view image data. Specifically, it improves the detection precision rate from 90.24% to 92.2%, the detection recall rate from 83.96% to 86.6%, and the mean average precision from 87.93% to 91.21%. Therefore, the improved SC4-YOLOv7 model not only ensures accurate detection but also enables efficient statistical analysis. It can serve as a valuable tool in future applications, such as detecting air conditioner external unit information in large-scale urban street view images and effectively responding to high-temperature and heat wave disasters.

Author Contributions: In this study, Z.T. contributed the manuscript conceptualization, data curation and analysis, manuscript writing, review and editing, funding acquisition; F.Y. made manuscript conceptualization, the data curation and analysis, manuscript writing, review and editing, figure plotting, and D.Q. made the data analysis and validation, review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported in parts by National Key R&D Program of China (Project No. 2022YFC3800700), and the National Natural Science Foundation of China (Grant No. 42171079).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McElroy, S.; Schwarz, L.; Green, H.; Corcos, I.; Guirguis, K.; Gershunov, A.; Benmarhnia, T. Defining heat waves and extreme heat events using sub-regional meteorological data to maximize benefits of early warning systems to population health. *Sci. Total Environ.* **2020**, *721*, 137678. [CrossRef] [PubMed]
2. Benmarhnia, T.; Deguen, S.; Kaufman, J.S.; Smargiassi, A. Vulnerability to heat-related mortality: A systematic review, meta-analysis, and meta-regression analysis. *Epidemiology* **2015**, *26*, 781–793. [CrossRef] [PubMed]
3. Russo, S.; Dosio, A.; Graversen, R.G.; Sillmann, J.; Carrao, H.; Dunbar, M.B.; Singleton, A.; Montagna, P.; Barbola, P.; Vogt, J.V. Magnitude of extreme heat waves in present climate and their projection in a warming world. *J. Geophys. Res. Atmos.* **2015**, *119*, 12500–12512. [CrossRef]
4. Takane, Y.; Ohashi, Y.; Grimmond, C.S.B.; Hara, M.; Kikegawa, Y. Asian megacity heat stress under future climate scenarios: Impact of air-conditioning feedback. *Environ. Res. Commun.* **2020**, *2*, 015004. [CrossRef]

5. Sera, F.; Gasparrini, A. Air conditioning and heat-related mortality: A multi-country longitudinal study. *Epidemiology* **2020**, *31*, 779–787. [CrossRef] [PubMed]
6. Romanello, M.; Mcgushin, A.; Di Napoli, C. The 2021 report of the lancet countdown on health and climate change: Code red for a healthy future. *Lancet* **2021**, *398*, 1619–1662. [CrossRef]
7. International Energy Agency. *Sustainable, Affordable Cooling Can Save Tens of Thousands of Lives Each Year*; IEA: Paris, France, 2023. Available online: https://www.iea.org/reports/sustainable-affordable-cooling-can-save-tens-of-thousands-of-lives-each-year (accessed on 8 October 2023).
8. Goel, R.; Garcia, L.M.T.; Goodman, A.; Johnson, R.; Aldred, R.; Murugesan, M.; Brage, S.; Bhalla, K.; Woodcock, J. Estimating city-level travel patterns using street imagery: A case study of using Google Street View in Britain. *PLoS ONE* **2018**, *13*, e0196521. [CrossRef]
9. Liu, X.; Song, Y.; Wu, K.; Wang, J.; Li, D.; Long, Y. Understanding urban China with open data. *Cities* **2015**, *47*, 53–61. [CrossRef]
10. Li, X.; Zhang, C.; Li, W.; Ricard, R.; Meng, Q.; Zhang, W. Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban For. Urban Green.* **2015**, *14*, 675–685. [CrossRef]
11. Zhang, F.; Wu, L.; Zhu, D.; Liu, Y. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. *ISPRS J. Photogramm. Remote Sens.* **2019**, *153*, 48–58. [CrossRef]
12. Hinton, G.E.; Salakhutdinov, R.R. Supporting online material for "Reducing the dimensionality of data with neural networks". *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
13. Szegedy, C.; Toshev, A.; Erhan, D. Deep Neural Networks for object detection. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1–9.
14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation tech report (v5). In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
16. Wang, C.; Hornauer, S.; Yu, S.X.; McKenna, F.; Law, K.H. Instance segmentation of soft-story buildings from street-view images with semiautomatic annotation. *Earthq. Eng. Struct. Dyn.* **2023**, *52*, 2520–2532. [CrossRef]
17. Yu, Q.; Wang, C.; McKenna, F.; Yu, S.X.; Taciroglu, E.; Cetiner, B.; Law, K.H. Rapid visual screening of soft-story buildings from street view images using deep learning classification. *Earthq. Eng. Eng. Vib.* **2020**, *19*, 827–838. [CrossRef]
18. Wang, C.; Antos, S.E.; Triveno, L.M. Automatic detection of unreinforced masonry buildings from street view images using deep learning-based image segmentation. *Autom. Constr.* **2021**, *132*, 103968. [CrossRef]
19. Guo, Y.; Wang, C.; Yu, S.X.; McKenna, F.; Law, K.H. AdaLN: A Vision Transformer for Multidomain Learning and Predisaster Building Information Extraction from Images. *J. Comput. Civ. Eng.* **2022**, *36*, 04022024. [CrossRef]
20. Fang, F.; Zeng, L.; Li, S.; Zheng, D.; Zhang, J.; Liu, Y.; Wan, B. Spatial context-aware method for urban land use classification using street view images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *192*, 1–12. [CrossRef]
21. Chen, B.; Feng, Q.; Niu, B.; Yan, F.; Gao, B.; Yang, J.; Gong, J.; Liu, J. Multi-modal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *109*, 102794. [CrossRef]
22. Gebru, T.; Krause, J.; Wang, Y.; Chen, D.; Deng, J.; Aiden, E.L.; Fei-Fei, L. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 13108–13113. [CrossRef]
23. Krylov, V.A.; Kenny, E.; Dahyot, R. Automatic discovery and geoagging of objects from street view imagery. *Remote Sens.* **2017**, *10*, 661. [CrossRef]
24. Chen, J.; Zhou, C.; Li, F. Quantifying the green view indicator for assessing urban greening quality: An analysis based on Internet-crawling street view data. *Ecol. Indic.* **2020**, *113*, 106192. [CrossRef]
25. Nguyen, Q.C.; Sajjadi, M.; McCullough, M.; Pham, M.; Nguyen, T.T.; Yu, W.; Meng, H.-W.; Wen, M.; Li, F.; Smith, K.R.; et al. Neighbourhood looking glass: 360º automated characterisation of the built environment for neighbourhood effects research. *J. Epidemiol. Community Health* **2018**, *72*, 260–266. [CrossRef] [PubMed]
26. Hu, L.; Wu, X.; Huang, J.; Peng, Y.; Liu, W. Investigation of clusters and injuries in pedestrian crashes using GIS in Changsha, China. *Saf. Sci.* **2020**, *127*, 104710. [CrossRef]
27. Yang, F.; Wang, M. Deep Learning-Based Method for Detection of External Air Conditioner Units from Street View Images. *Remote Sens.* **2021**, *13*, 3691. [CrossRef]
28. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. *arXiv* **2022**, arXiv:2207.02696.
29. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021.
30. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
31. Jiang, K.; Xie, T.; Yan, R.; Wen, X.; Li, D.; Jiang, H.; Jiang, N.; Feng, L.; Duan, X.; Wang, J. An Attention Mechanism-Improved YOLOv7 Object Detection Algorithm for Hemp Duck Count Estimation. *Agriculture* **2022**, *12*, 1659. [CrossRef]
32. Bochkovskiy, A.; Wang, C.Y.; Liao HY, M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.

33. Ni, Y.; Mao, J.; Fu, Y.; Wang, H.; Zong, H.; Luo, K. Damage Detection and Localization of Bridge Deck Pavement Based on Deep Learning. *Sensors* **2023**, *23*, 5138. [CrossRef]
34. Ni, Y.; Mao, J.; Wang, H.; Fu, Y.; Xi, Z. Corroded and loosened bolt detection of steel bolted joints based on improved you only look once network and line segment detector. *Smart Struct. Syst.* **2023**, *32*, 23–35. [CrossRef]
35. Zhang, X.; Zeng, H.; Guo, S.; Zhang, L. Efficient Long-Range Attention Network for Image Super-Resolution. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 649–667.
36. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
37. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020. [CrossRef]
38. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
39. Zhang, F.; Zhou, B.; Liu, L.; Liu, Y.; Fung, H.H.; Lin, H.; Ratti, C. Measuring human perceptions of a large-scale urban region using machine learning. *Landsc. Urban Plan.* **2018**, *180*, 148–160. [CrossRef]
40. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 13–19 June 2020.
41. Van Etten, A. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv* **2018**, arXiv:1805.09512.
42. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
43. Wang, C.; Bai, X.; Wang, S.; Zhou, J.; Ren, P. Multiscale Visual Attention Networks for Object Detection in VHR Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 310–314. [CrossRef]
44. Ferrari, L.; Rosi, A.; Mamei, M.; Zambonelli, F. Extracting urban patterns from location-based social networks. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, IL, USA, 1 November 2011; pp. 9–16.
45. Yu, L.; Zhengwei, S.; Chaogui, K.; Yong, G.; Peter, C. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS ONE* **2014**, *9*, e86026.
46. Liu, Y.; Kang, C.; Gao, S.; Xiao, Y.; Tian, Y. Understanding intra-urban trip patterns from taxi trajectory data. *J. Geogr. Syst.* **2012**, *14*, 463–483. [CrossRef]
47. Hoffmann, E.J.; Wang, Y.; Werner, M.; Kang, J.; Zhu, X.X. Model Fusion for Building Type Classification from Aerial and Street View Images. *Remote Sens.* **2019**, *11*, 1259. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.