Article

VOD: Vision-Based Building Energy Data Outlier Detection

Jinzhao Tian [1](ID), Tianya Zhao [2,†](ID), Zhuorui Li [3,†](ID), Tian Li [1](ID), Haipei Bie [1](ID) and Vivian Loftness [1,*](ID)

1 School of Architecture, Carnegie Mellon University, Pittsburgh, PA 15213, USA; jinzhaot@andrew.cmu.edu (J.T.); tili@cmu.edu (T.L.); hbie@andrew.cmu.edu (H.B.)
2 Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA; tzhao010@fiu.edu
3 School of Engineering, The University of Kansas, Lawrence, KS 66045, USA; zhuorui@ku.edu
* Correspondence: loftness@cmu.edu; Tel.: +1-412-983-6055
† Work performed while at Carnegie Mellon University.

Abstract: Outlier detection plays a critical role in building operation optimization and data quality maintenance. However, existing methods often struggle with the complexity and variability of building energy data, leading to poorly generalized and explainable results. To address the gap, this study introduces a novel Vision-based Outlier Detection (VOD) approach, leveraging computer vision models to spot outliers in the building energy records. The models are trained to identify outliers by analyzing the load shapes in 2D time series plots derived from the energy data. The VOD approach is tested on four years of workday time-series electricity consumption data from 290 commercial buildings in the United States. Two distinct models are developed for different usage purposes, namely a classification model for broad-level outlier detection and an object detection model for the demands of precise pinpointing of outliers. The classification model is also interpreted via Grad-CAM to enhance its usage reliability. The classification model achieves an F1 score of 0.88, and the object detection model achieves an Average Precision (AP) of 0.84. VOD is a very efficient path to identifying energy consumption outliers in building operations, paving the way for the enhancement of building energy data quality, operation efficiency, and energy savings.

Keywords: AI-driven; deep learning; outlier detection; load shape; building energy

1. Introduction

Building energy consumption accounts for 30% of total energy use in the world and almost 30% of total carbon emissions according to IEA 2022 [1]. To comply with the 2050 Paris goal of net zero emissions, it is essential to reduce building energy use while maintaining normal building operations [2]. To improve building equipment effectiveness while reducing energy consumption, rapidly detecting energy consumption outliers becomes essential. Energy use abnormalities often manifest in an irregular pattern, such as point outliers, contextual outliers, and collective outliers, representing an ineffectiveness or failure in equipment operation or sensors. Failing to identify these failures can lead to a failure to deliver critical building services, as well as higher energy waste and carbon emissions [3]. Outliers occur when sensors in the buildings become faulty or the building operation requires optimization or commissioning [4–7]. Outliers should be identified and removed before data analysis, as they may adversely impact the data quality, leading to degraded analysis results or performance [8,9]. With advances in machine learning and the increased use of smart meters and building automation systems in commercial buildings, unwanted energy consumption detection has become more feasible and effective. Energy consumption is logged per set time interval, usually from one minute to an hour. A database containing energy consumption against time is recorded as "time-series" data, the core of abnormal energy use detection. Many outlier detection methods utilize machine learning to spot issues in energy consumption data. However, the following three challenges remain:

First, the lack of labeled, high-quality data limits researchers' options. Himeur et al. [10] stated that there are very few data that are labeled, and the amount of outliers in the existing dataset is very limited. The lack of annotated data is caused by the difficulties in finding them, and it is very expensive and labor-intensive to label these outliers. Even if a dataset is properly labeled, such an unbalanced dataset makes it not suitable to train an outlier detection model. In the end, the amount of normal data has greatly surpassed the amount of outliers, leading to a biased model that cannot accurately identify faults. To combat this, it is important to broadly label and balance the normal data and outliers and generate more sets of outliers for research.

Second, there is a significant variation in outliers in real-world data. While outliers can be theoretically defined and categorized, real-world data present a more complex picture. Outliers can emerge due to a variety of reasons, including meter connection issues and various abnormal consumption behaviors [10]. Consequently, these outliers display a wide range of characteristics in terms of magnitude, duration, frequency of occurrence, and patterns. Such diversity poses a considerable challenge to conventional machine learning methods. This diversity in outlier attributes necessitates a more generalized approach in outlier detection methodologies, which can adapt to the wide range of variations and accurately identify anomalies regardless of their distinct features.

Third, many outlier detection methods favor using deep learning algorithms with "black-box" models due to their superior prediction performance. Using "black-box" models poses difficulties in explaining the reason for the output. The opaqueness in model functionality presents challenges in elucidating the rationale behind their outputs, leading to a trust issue among users without specialized knowledge in this domain. For example, Copiaco et al. [11] proposed image-based outlier detection using deep learning. Despite achieving high performance, the model is a "black box", meaning the output cannot be explained, and its decision-making processes are uninterpretable. Therefore, a clear explanation of the model is required to bridge the gap between model performance and user trust.
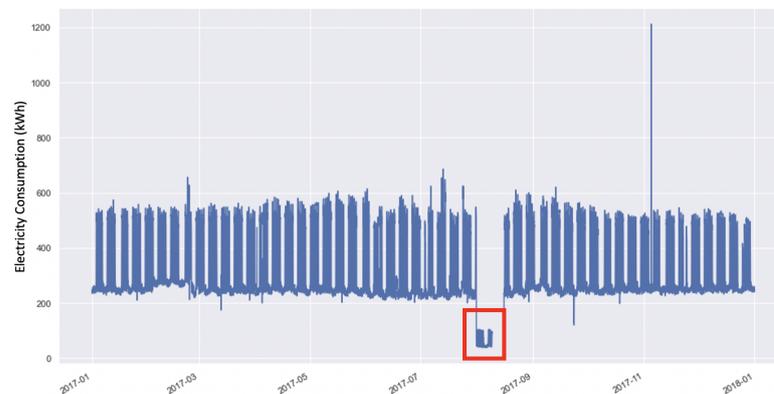
2. Literature Review

2.1. Conventional Outlier Detection Methods

The "Interquartile Range" (IQR) method is commonly used to detect outliers for its simplicity. The method calculates outliers by setting boundaries at the 25th percentile minus 1.5 times the IQR and the 75th percentile plus 1.5 times the IQR. Theoretically, it is effective for identifying extreme outliers and cleaning the data. However, in the real world, the boundaries are not optimal, as illustrated in Figure 1, which is a one-year weekly plot of the electricity consumption data in one building; the upper boundary defined by the IQR method detects the normal electricity consumption data as outliers, while the lower boundary is too low to detect the abnormal pattern in the morning and evening.

To overcome the issue, a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method [12] becomes more commonly used in detecting outliers in building energy data [13–16]. The algorithm clusters the data points into recognizable groups when they are close to each other. If no group is assigned to a data point, it will be defined as an outlier. It operates based on the following two key parameters: a measure of distance and a minimum count of points needed to establish a dense region. However, if the minimum number of points is not defined correctly, when a substantial group of outliers deviates from the usual pattern, as shown in Figure 2, the DBSCAN method may struggle to accurately identify the extreme values. To effectively detect outliers in building energy data, it is critical to have a comprehensive understanding of the typical energy usage pattern.

Figure 1. An example to illustrate the problem of the "IQR" method. The red circled parts highlight the "False Positive" outliers detected by the upper boundary and the "False Negative" outliers not detected by the lower boundary.

Figure 2. An example to illustrate the problem of the DBSCAN method. The red rectangle part highlights the "group outliers" which are hard to detect by the DBSCAN method.

2.2. Outlier Detection Based on Load Shape

"Whole-building electric load" represents the total electrical power consumed by a building at a given moment, which can change based on demand changes in lighting, HVAC, and plug-load device usage. Analyzing the load shape over time can provide valuable insights, including energy waste, equipment issues, and HVAC operation problems [17]. Previous studies have utilized different load shape parameters to extract useful features from the load shape. Luo et al. [18] used three parameters, namely the peak-based load ratio, workday/non-workday load ratio, and one-hour duration, to interpret the load shape. Liu et al. [19] divided the daily profiles into four segmentations representing off-time, rise time, daytime, and evening based on the regular occupancy schedule of office buildings and added the daily peak-to-valley difference to the clustering analysis. However, these methods often require manual threshold settings, which can introduce bias, limiting their ability to fully represent the dataset.

To overcome these limitations, researchers have turned to deep learning, given its capability as a universal function approximator. Fan et al. [20] employed an auto-encoder to learn the normal energy usage patterns of buildings and detect outliers. Zheng et al. [21] converted 1D time-series electricity consumption into a 2D array by aligning the consumption data of several weeks together. However, each of these methods struggles to capture the temporal and spatial correlations of the time-series data due to the representational limitation of the 1D time-series data. To address these issues, Fahim et al. [22] and Copiaco et al. [11] innovated by converting building energy data into 2D plots, and Fahim et al. [22] transformed the energy data into a 2D Markov transition field, while Copiaco et al. [11] combined energy data with occupancy data and other features extracted from the energy

data to create a 2D array. Then, they converted the 2D array to a 2D color map image and used pre-trained Convolutional Neural Network (CNN) models to detect the outliers. However, all these methods can only classify whether outliers occur in a detected window and are unable to localize the occurrence of outliers precisely.

2.3. Supervised Outlier Detection

Current deep learning models for outlier detection techniques are mainly unsupervised models. Unlike unsupervised models, supervised models can provide a clearer metric for performance evaluation, since they use labeled data, allowing for precise measurement of accuracy [23]. Recently, studies have also shown high efficiency for supervised outlier detection, such as the studies by Bawono and Bachtiar [24], Paulheim and Meusel [25], and Aggarwal and Aggarwal [26]. The differences and comparisons between unsupervised and supervised models regarding building science data outlier detection methods were discussed by [27]. Compared to unsupervised learning models, supervised learning models are not that widely applied to building science.

Bawono and Bachtiar [24] addressed the challenges of using supervised learning for outlier detection, particularly the issue of imbalanced data classification due to the typically small proportion of outliers in datasets. The authors showed a promising result to resolve the current issues of supervised learning-based outlier detection. Araya et al. [28] proposed a new framework based on collective contextual anomaly detection using a sliding window (CCAD-SW) to identify unusual energy use in smart buildings, which was further enhanced by the Ensemble Anomaly Detection (EAD) framework combining various classifiers. Using real-world data, the EAD was shown to increase the sensitivity of CCAD-SW by 3.6% and decrease false alarms by 2.7%, proving effective in energy monitoring. Shoemaker and Hall [29] conducted a study using an ensemble voting method in supervised learning for anomaly detection. Their approach combined random forests with distance-based outlier partitioning. The research found that this method yielded accuracy results comparable to those of the same techniques without the partitioning element. This demonstrated the ensemble voting method's effectiveness in maintaining accuracy while incorporating additional outlier partitioning strategies. Miyata et al. [30] mentioned that supervised model outlier detection is designed to classify unseen data as either normal or an outlier. Beyond just detecting outliers, it can also determine the specific type of outlier. This feature is particularly valuable in applications like HVAC system Fault Detection and Diagnostics (FDD), where understanding the cause of the outlier is crucial for effective problem-solving and maintenance. Xu and Chen [31] proposed an anomaly detection method for Ground Source Heat Pump (GSHP) systems using long short-term memory (LSTM) and Grubbs' test. It predicts energy consumption, identifies three types of operational anomalies, and validates them through field checks and expert opinions, enhancing GSHP system efficiency and operation.

The current outlier detection by supervised learning models has some advantages over unsupervised learning models, including parameter selection, potential accuracy, and the capacity to control. However, most of the current supervised outlier detection methods for building smart meter datasets focus on numeric datasets, creating potential interpretation issues for building owners and managers. These existing supervised learning approaches are not adequate for building energy load shape outlier detection.

2.4. Model's Explainability

AI-driven outlier detection techniques and unsupervised or supervised learning algorithms include three categories, which are "white-box", "gray-box", and "black-box" models. Their advantages and disadvantages have been discussed regarding the concept of "explainability" for transparency to the users [2,32,33]. White-box models, known for their interpretability, allow for a clear understanding of how input data lead to specific conclusions, making them advantageous in contexts where transparency and explainability are critical, such as linear, polynomial, and ridge regressions. However, they often struggle

to achieve the same level of accuracy as black-box models, especially with complex data. On the other hand, black-box models are typically more accurate and can handle complex and high-dimensional data effectively, i.e., neural network-based deep learning models. The downside is their lack of interpretability, as the internal workings are not easily understood, making them less suitable for situations where understanding the decision-making process is important [34]. Gray-box models, such as tree-boosting models, have better interpretation but still include many aspects that cannot be explained [35]. Additionally, compared to the black-box models, especially when dealing with time-series data types, "gray-box" models underperform [36].
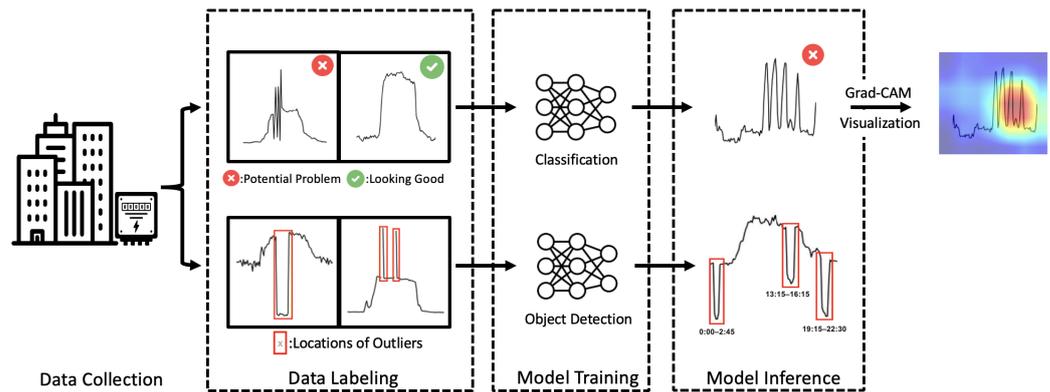
Multiple studies explored different practices to overcome the explainability of deep learning models for building science and other study areas. Machlev et al. [37] discussed the challenge of understanding "black-box" models, especially in the power systems field, where accountability is crucial. They introduced Explainable Artificial Intelligence (XAI) techniques as a solution to improve the explainability of these models, aiming to make their outputs more comprehensible. They reviewed common challenges, recent works, and ongoing trends in XAI for power system applications, aiming to inspire further research and discussions on this emerging topic. Somu et al. [38] introduced a CNN-LSTM deep learning framework for the prediction of building energy consumption, combining k-means clustering, convolutional neural networks (CNNs), and LSTM networks. Tested on real data from a building in IIT-Bombay, India, CNN-LSTM showed superior performance in forecasting by effectively capturing spatiotemporal patterns in energy data compared to state-of-the-art models. The results also showed energy prediction comparisons as graphical illustrations. Li et al. [39] presented a deep learning approach for the prediction of building energy consumption, combining stacked autoencoders (SAEs) with an extreme learning machine (ELM) to enhance accuracy. SAEs extracted features, while the ELM predicted energy use. The study also plotted energy prediction results to better visualize model performance. Similarly, Fan et al. [40] proposed deep learning-based methods for building cooling load prediction. The study also found that the feature engineering process of an unsupervised learning model can improve prediction performance.

While current studies extend their interest to deep learning visualization and explainability, most of studies focus on the visualization of the results or comparisons instead of the model itself [10]. This leaves a gap for decision-makers to understand the analysis mechanism inside of the model.

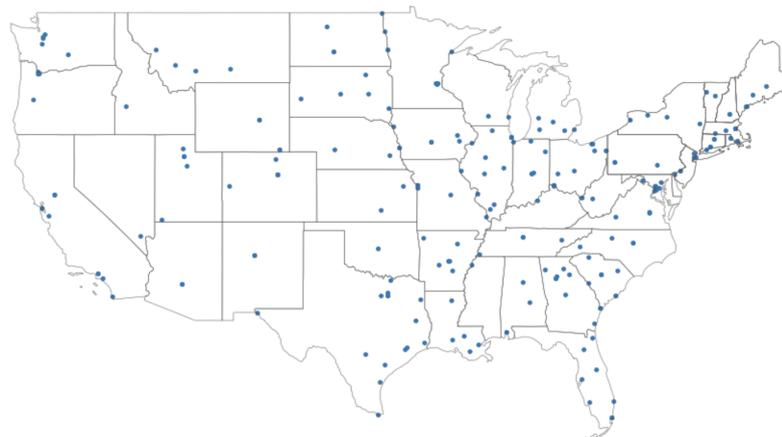## 3. VOD Outlier Detection Methodology

3.1. Overview

Figure 3 illustrates the VOD outlier detection methodology. The process starts with the collection and transformation of the building energy consumption data into daily profile time-series plots. Subsequently, the following two distinct models are developed to meet different usage purposes: a classification model for broad-level outlier detection, which identifies the presence of outliers in the building energy consumption, and an object detection model that precisely pinpoints the timing of these outliers. Additionally, as the classification model does not provide the location information of the outliers, Grad-CAM is applied to visually interpret the outputs of the model.

Figure 3. Flowchart of the VOD outlier detection approach. A demo of the object detection model can be found VOD Object Detection Demo (accessed on 22 April 2024).

3.2. Dataset

This study utilizes a comprehensive dataset covering 290 buildings during the 2016–2019 period sourced by the U.S. General Services Administration (GSA), an independent agency of the United States for the management and support of the basic functioning of federal agencies. The dataset includes detailed 15-minute-interval records of electricity consumption for each of the 290 buildings, capturing variations in their electricity usage over time. To refine the focus on outlier detection during regular workdays, data corresponding to weekends and public holidays were excluded from the analysis (Figure 4).
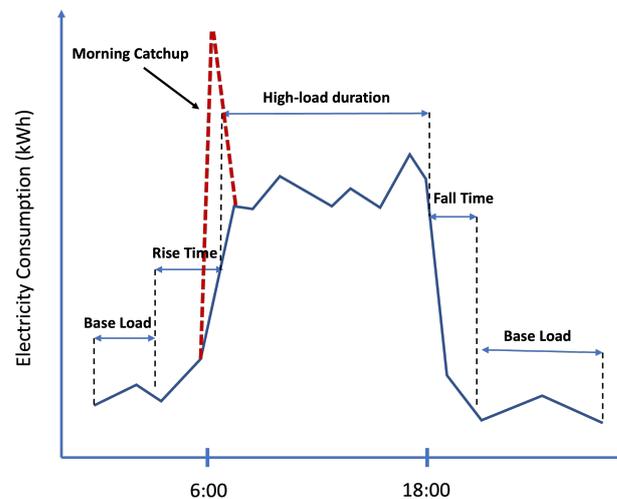

Figure 4. Map of the Buildings in the dataset. The blue points indicate the locations of the buildings.

3.3. Typical Daily Electricity Consumption Load Shape of Office Buildings on Workdays

The shape of electricity consumption in office buildings during the workday can be divided into the following five parts: morning near base load, load rise time, high-load duration, load fall time, and evening near base load (see Figure 5) [18]. It begins with the morning near base load phase before 6 AM, characterized by minimal energy usage due to low occupancy and limited operational activities, maintaining only essential systems such as security and safety lighting. This is followed by the "load rise time", where electricity usage gradually increases as the building prepares for the workday—lights are turned on, HVAC systems ramp up, and office equipment is activated. During this period, a sudden spike, known as "morning catch-up" may occur due to the HVAC system turning on to rapidly pre-condition spaces [17]. This time period is followed by a normal "high-load duration", which coincides with standard office hours, where sustained energy consumption is at its maximum due to the full-scale operation of all systems and high occupancy. After work hours, during the "fall time" load phase, there is a noticeable decline in consumption as employees leave and lights and equipment are turned off, although

some systems, such as HVAC, remain partially active. Finally, the "evening near base load" phase posts around 8 PM, where the building returns to a low-energy state, similar to the early morning, operating under minimal activity until the next day.
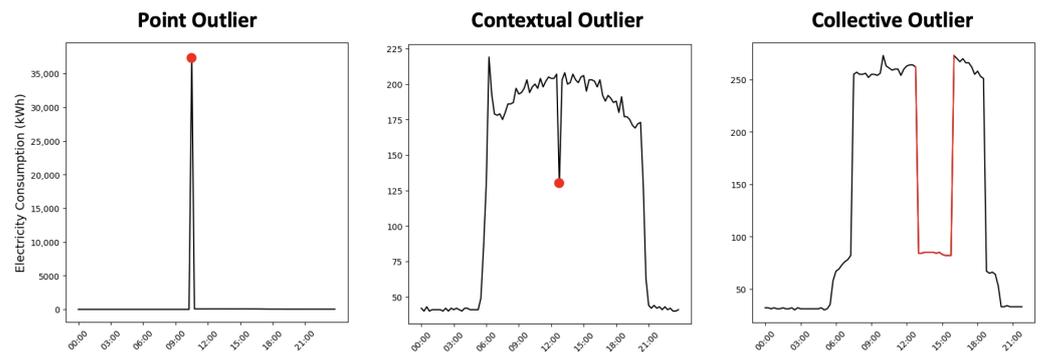

Figure 5. Typical daily electricity consumption shape of the office building during a workday.

3.4. Definition of Outliers in Daily Electricity Consumption

Hawkins described an outlier as an observation that "deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism" [41]. Based on this definition, outliers can be further classified into the following three categories: point, contextual, and collective outliers [3,42].

- Point outliers refer to data instances that are significantly different from the majority of data points in a dataset.
- Contextual outliers, also called conditional outliers, are anomalous instances in a specific context. They usually have relatively larger or smaller values with respect to their adjacent values. However, when viewed independently, they will fall within the normal range expected for the signal.
- Collective outliers, also known as group outliers, are a series of data points that are anomalous with respect to the entire data set. They usually show an unusual shape compared with the entire dataset.

In the context of daily electricity consumption during workdays, point outliers are those instances that significantly deviate from the normal range of records within the workday. On the other hand, contextual outliers are within the normal electricity consumption range, but if they are considered with adjacent records, they will deviate from the normal pattern. As illustrated in Figure 6, a contextual outlier, though within the standard range of energy consumption, reveals itself as an abrupt variation—a sudden drop or spike—that deviates from the typical "workday shape" when analyzed with adjacent records. Collective outliers, meanwhile, represent a series of missing records or records that, together, form an abnormal pattern within the daily workday profile. They are manifested as gaps or irregular fluctuations, such as bumps and depressions.

Figure 6. Three Types of Outliers in Daily Office Electricity Consumption.
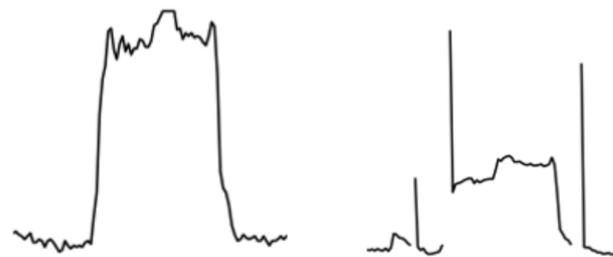
3.5. Classification Dataset

To facilitate analysis and classification, the daily energy consumption profiles for each building were transformed into binary images. The dataset consists of a total of 15,640 images, with each image standardized to a size of 224 pixels by 224 pixels. This image size was chosen due to its prevalence in image classification tasks.

To mitigate potential biases in the dataset, a shuffling process was performed prior to image conversion. This approach ensures that a diverse representation of buildings' electricity consumption patterns is preserved, enhancing the dataset's capacity to generalize across different scenarios.

Each binary image was manually labeled based on its corresponding daily electricity consumption pattern. Two distinct categories were defined, namely "looking good" and "potential problems", based on whether any outliers could be observed or not.

After labeling, the dataset was further divided into the following three subsets: a training set, a validation set, and a test set. The distribution ratio was set at 8:1:1, respectively. This division ensures that the model is trained on a substantial portion of the dataset while retaining independent subsets for validation and final performance evaluation (Figure 7).

Figure 7. Examples of the classification dataset. The left panel adheres to the typical daily electricity consumption pattern during workdays as previously defined; therefore, it is labeled as "looking good". In contrast, the right panel contains three spikes (the left spike is a contextual outlier, and the remaining two are point outliers) in the daily electricity consumption records. Due to these outliers, it is labeled as "potential problems".
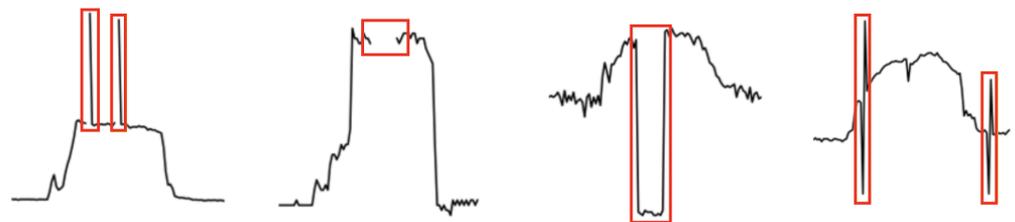
3.6. Object Detection Dataset

To delve deeper into identifying potential problems within the "potential problems" category, a subset of 2160 images was randomly selected. This subset is intended to focus on instances where potential outliers in electricity consumption occur. In order to facilitate object detection, bounding boxes were manually placed around the areas of interest within these images. These bounding boxes indicate regions where consumption patterns deviate from the typical daily electricity consumption shape.

The object detection dataset was then shuffled to ensure a representative mix of images during the training, validation, and testing phases. This shuffling process contributes to

reducing any inherent bias that may be present in the dataset, thus promoting the model's capacity to generalize effectively.

Similar to the classification dataset, the dataset was divided into training, validation, and test sets with a ratio of 8:1:1. To facilitate compatibility with the object detection model, all images within the dataset were resized to uniform dimensions of 320 pixels by 320 pixels. This standardization ensures that the input dimensions conform to the expectations of the model architecture, allowing for the integration of the dataset into the object detection pipeline (Figure 8).

Figure 8. Examples of the object detection dataset. The coordinates of the top-left and bottom-right corners are recorded in the labeled dataset to define the bounding boxes (the red boxes in the figure) of the outliers.
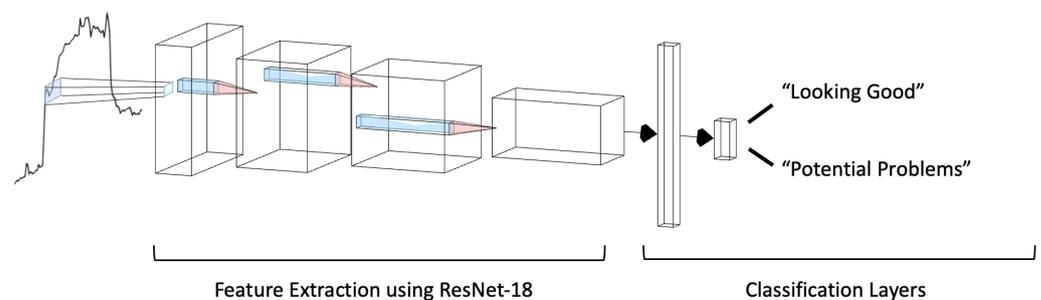
3.7. Model Development

3.7.1. Classification Model

By converting the daily electricity consumption data into 2D image representations, powerful and well-established deep learning models from the field of computer vision can be leveraged for outlier detection. Convolutional neural networks (CNNs) have become one of the most widely deployed network architectures across numerous computer vision tasks [43]. Taking advantage of their capabilities, this paper applies CNNs to the electricity consumption images to identify outliers.

A residual network (ResNet) [44] is a classic CNN architecture that has significantly impacted the field of computer vision and deep learning [45]. The key innovation behind ResNet is the introduction of residual connections, also known as skip connections or shortcut connections. These connections allow for the training of very deep neural networks without suffering from the vanishing gradient problem. As a result, ResNet excels at learning hierarchical abstractions and nuanced patterns in data through their exceptionally deep representations.

This paper deploys ResNet-18, which is a specific variant of the ResNet family. An overview of workday daily electricity classification is shown in Figure 9. Feature extraction analyzes the 2D electricity images and extracts meaningful patterns from them using residual convolutional blocks. Then, a fixed linear layer is applied to predict whether the electricity profile looks good or has potential problems.
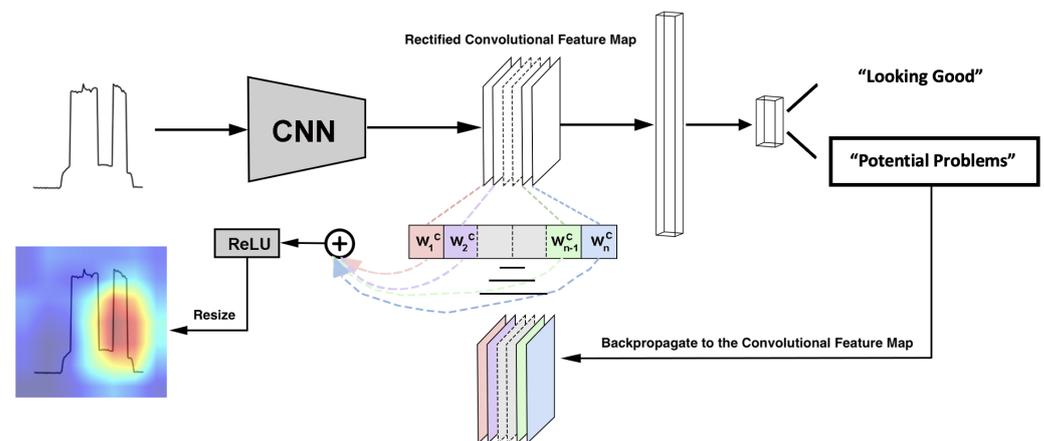

Figure 9. Flowchart of workday daily electricity outlier classification.

3.7.2. Classification Model Visual Explanation via Grad-CAM

Grad-CAM is an important method to address the "black-box" nature of deep learning models and improve their interpretability [46]. As people understand the model more deeply, users will have more trust and confidence in deploying these models.

The foundational idea behind Grad-CAM is to highlight the areas of input that are more crucial for the model's decision-making. As shown in Figure 10, the input daily energy consumption image passes through the CNN, and the network returns the prediction of "potential problems," indicating an outlier in the given daily electricity consumption data.

The last convolutional layer of the CNN model is often chosen for Grad-CAM because it provides the most informative combination of high-level feature representation and spatial context, which is essential for creating meaningful and interpretable visual explanations of the model's decision-making process. To highlight the area that contributes most to the outlier, the gradients of the feature maps related to the output of "potential problems" are calculated and averaged as the weight for each feature map in the last convolutional layer. After that, the aggregated weighted feature map is passed through the ReLU active function to remove the non-positive weights, which contribute to the "looking good" class instead of the outlier. Finally, the feature map is converted into a heatmap and resized to the same size as the input image to highlight the area of the most informative part for the class of "potential problems" within the input daily electricity consumption plot.

Figure 10. Flowchart of the classification model visual explanation via Grad-CAM [46]. (Rectified convolutional feature map: the convolution layer after the ReLU active function).

3.7.3. Object Detection Model

YOLOv5 [47] is the most popular one-stage object detector and is renowned for its widespread application across various fields [48–50]. YOLOv5 has four different models; in this study, YOLOv5s (small) is used. This choice is driven by the relative simplicity of the task because only one class—"outlier"—needs to be detected. As shown in Figure 11, the model consists of the following three primary components: the backbone, which is used to extract the feature representation of the input images; the neck, which is used to combine the image features in different layers in the backbone for further prediction; and the head, which is responsible for generating the final detection, including bounding boxes, confidence scores, and class predictions. Moreover, YOLOv5 uses a Path Aggregation Network (PANet) in the neck part to help the network extract and aggregate features better, leading to improved performance in object detection tasks [51].
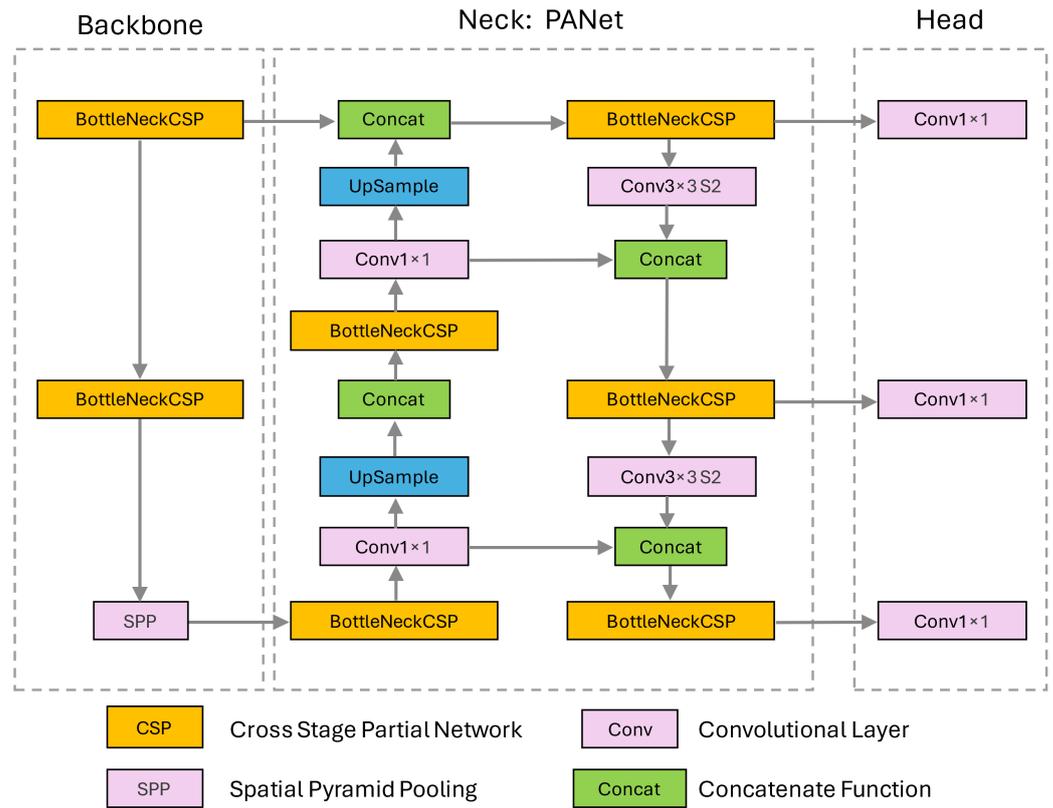
Figure 11. Flowchart of YOLOv5 [52].

3.8. Loss Function and Evaluation Metrics

Different loss functions and evaluation metrics are used in training and evaluating the two models—classification and object detection. For the classification model, Binary Cross-Entropy loss (BCELoss) is used for the binary classification task. For the object detection model, the loss function is a combination of three components, namely the bounding box loss (l_{bbox}), the object loss (l_{obj}), and the classification loss (l_{cls}). The l_{bbox} is used to ensure the accuracy of the bounding-box predictions, while the l_{obj} assesses the model's confidence in the presence of an object within each bounding box, focusing on the model's ability to discern whether a bounding box indeed contains an object accurately. Lastly, the l_{cls} is designed to ensure the class prediction accuracy by assessing the difference between predicted class probabilities and the actual class labels.

Regarding the evaluation metrics, the classification model is assessed with precision, recall, and the F1 score. These metrics provide a holistic view of the model's performance, accounting for both the accuracy of positive predictions and the model's ability to detect all relevant instances. For the object detection model, the average precision (AP) offers a nuanced evaluation by considering the precision–recall curve across various thresholds, thereby encapsulating the model's accuracy and robustness in object detection tasks. Specifically, this study focuses on AP_0.5, a variant of AP calculated at an Intersection over Union (IoU) threshold of 0.5. This threshold is a standard metric in many object detection challenges and benchmarks [53,54]. It ensures that models are not overly penalized for slight inaccuracies in bounding-box predictions while maintaining a reasonable precision standard.

$$\text{BCELoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\log(p_i) + (1 - y_i)\log(1 - p_i))] \tag{1}$$

$$\text{Object Detection Loss} = l_{bbox} + l_{obj} + l_{cls} \tag{2}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4}$$

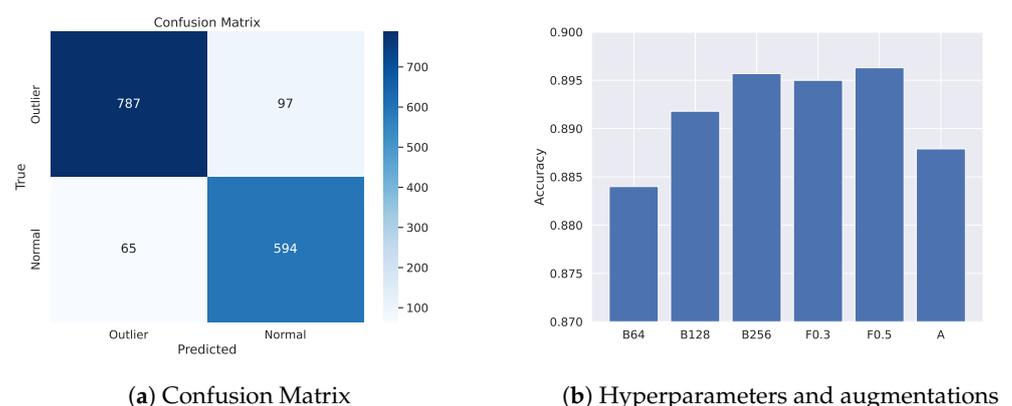$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{AP} = \sum_{i=1}^{n} (R_i - R_{i-1}) P_i \tag{6}$$

N the number of observations;
y_i the ground truth of the ith observation, which can be 0 or 1;
p_i the predicted probability of the ith observation being of class 1;
l_{bbox} the bounding-box prediction loss;
l_{obj} the object loss;
l_{cls} the classification loss;
R_i the recall at the ith IoU threshold;
P_i the precision at the ith IoU threshold.

4. Outlier Detection Success and Discussion

4.1. Outlier Classification

A confusion matrix summarizing the performance of our outlier classification model is presented in Figure 12a. ResNet-18 achieves a precision of 85.96% and a recall of 90.13%, resulting in an F1 score of 0.88. These positive results demonstrate the feasibility and efficacy of utilizing image-based features for the detection of outliers in building energy consumption data. Furthermore, the impacts of varying hyperparameters and augmentations are examined. Figure 12b summarizes the overall accuracy with different batch sizes and augmentations. In this figure, "B" denotes the batch size during training, "F" indicates that random horizontal flipping is applied to input images, and "A" specifies the use of affine transformations as an augmentation strategy. It is worth noting that employing a larger batch size yields superior results. This improvement may be attributed to the fact that a greater variety of outliers is encountered simultaneously during the training process. Specifically, when using a flip probability of 0.5 for data augmentation, the accuracy rises to 89.63%. However, other data augmentation techniques do not further improve accuracy.

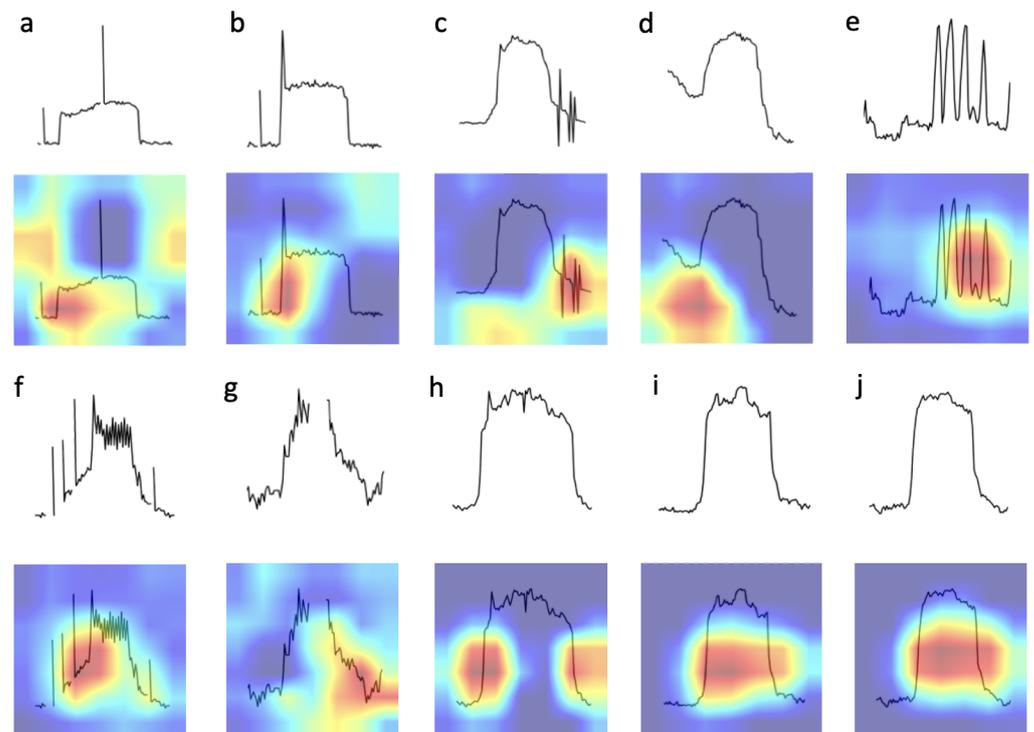(**a**) Confusion Matrix   (**b**) Hyperparameters and augmentations

Figure 12. Outlier classification results.

4.2. Adding Grad-CAM to Visualize the Classification Model

The visualization of the Outlier Classification model is achieved using Grad-CAM, as illustrated in Figure 13. The figure presents ten unique time-series images labeled from a to j. For each image, the upper portion displays the input daily electricity consumption plot, while the lower portion shows the Grad-CAM visualization. The heatmap in the Grad-CAM results highlights the areas critical for the classification model's predictions.

Images a to g are classified in the "potential problems" category, whereas images h to j correspond to the "looking good" category. For images a to g, the Grad-CAM visualization emphasizes the area with important outlier features learned from the model. Conversely, for images h to j, the heatmap highlights the areas representing the normal patterns in daily electricity consumption.

Figure 13. Visualization of the classification model via Grad-CAM. Images (**a–g**) are classified as "potential problems", while images (**h–j**) are classified as "looking good". The red regions in the heatmaps indicate the most informative parts used by the model for making classification decisions.

For the results from the "looking good" category, (h–j), the highlighted areas are the period of "rise time" and "fall time." This indicates that the classification model recognizes the normal daily electricity usage patterns by finding the intervals of electricity load increase and decrease correlating with the start and end of working hours, respectively. Moreover, the model also expects the normal daily electricity usage to be relatively low during the morning and evening base load periods. For example, in image d, the heatmap from Grad-CAM marks the abnormal patterns observed in the morning. It reveals a group of "collective outliers" where the morning load surpasses the evening load. This difference indicates unusual electricity usage since the previous evening, leading to its classification under "potential problems". Further examples illustrating the key regions for the "looking good" category are detailed in Figure A1.
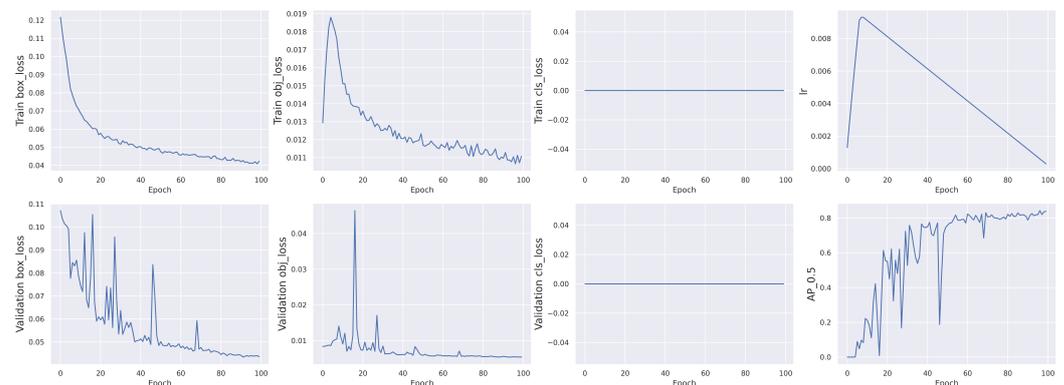
The "potential problems" category exhibits significantly more variability and complexity compared to the "looking good" category. This is due to the diversity in the load shape; outliers appear sporadically, varying in time, magnitude, and duration. Unlike the relatively simple "rise and fall" pattern, here, the classification model must identify hidden features within a broad daily load shape that includes these outliers. Images a to g highlight the regions crucial for the classification model to predict the "potential problems" category. While the model is capable of detecting different types of outliers, its decision-making may be compromised when multiple outliers occur in a day's energy consumption. This limitation underscores the need for an object detection model that can more effectively discern and react to these irregularities. Additional examples are illustrated in Figure A2.

4.3. Outlier Object Detection

The hyperparameter settings for training of the outlier object detection model are shown in Table 1. The four initial parameters are fundamental to the training, while the latter four relate to the probability of data augmentations applied during training. Figure 14 illustrates the training process of the model. After 100 epochs of training, the model achieves AP_0.5 of 0.84, indicating its effectiveness in identifying and pinpointing outliers in daily energy consumption data. The l_{bbox} reduces to 0.042 and 0.044 for the training and validation sets, respectively. Similarly, the l_{obj} decreases to 0.011 for training and 0.005 for validation. The l_{cls} remains at 0, since there is only one object class, and the model does not need to classify different classes for the detected objects. The minimal discrepancies between the training and validation results reflect the model's robust generalization capabilities and the absence of overfitting.

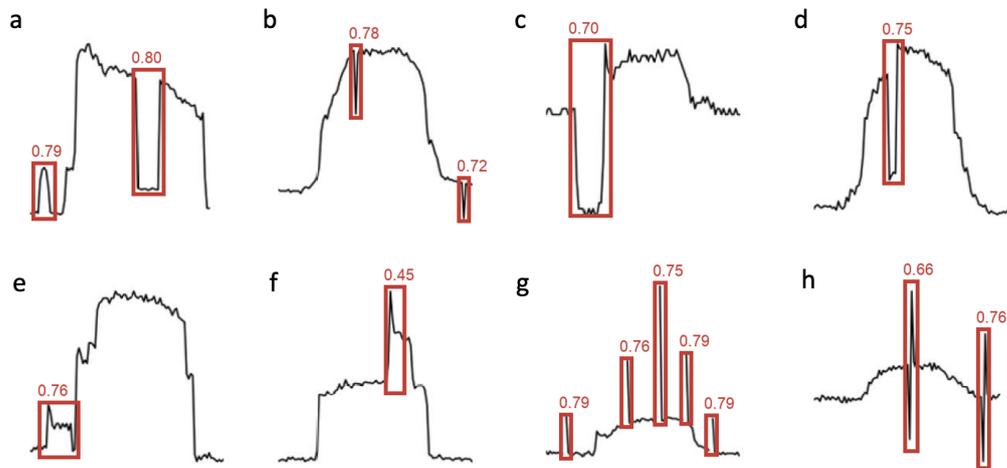**Table 1.** Hyperparameter Settings of the object detection model.

Hyperparameter	Settings
lr	0.01
epochs	100
batch_size	128
warm_up epochs	3
scale	0.5 (probability)
mosaic	1.0 (probability)
translate	0.1 (probability)
horizontal flip	0.5 (probability)

Figure 14. The training process of the outlier object detection model.

Figure 15 illustrates the model's performance in detecting various types of outliers in daily energy consumption data. It successfully detects the point outliers in examples g and h; the contextual outliers in example b; and the collective outliers in a, c, d, e, and f. In images b, g, and h, the model identifies the point and contextual outliers that occur regularly in a single day, indicating potential sensor or connection errors with the smart meters. These errors could lead to significant impacts, such as inaccurate billing and compromised energy distribution. In images a, c, d, e, and f, the detected collective outliers highlight the abnormal electricity consumption patterns. The outliers in a, c, and d indicate significant drops in electricity usage, suggesting the abnormal shutdown of the electricity devices, which could cause disruptions in the normal operation of the buildings. For the outliers in a, e, and f, the model pinpoints the abnormal increases in the electricity load, which may be due to unauthorized consumption or faulty equipment and will lead to increased energy costs and potential overloading of the power grid. Notably, in examples a, b, g, and h, the model faces the challenge of multiple outliers occurring simultaneously within a single day, yet it successfully discerns and localizes each outlier. The results
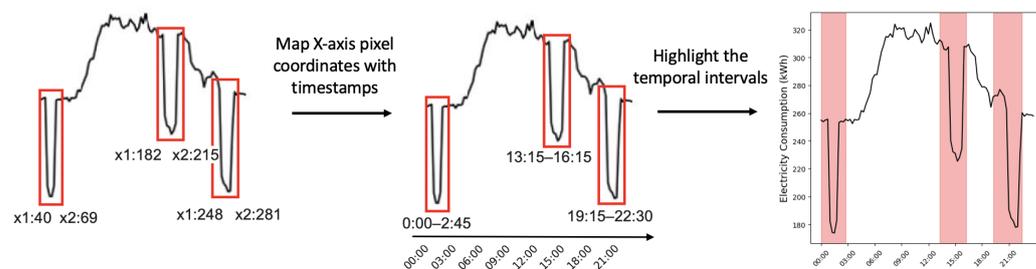
underscore the model's ability to manage and discern between multiple and distinct outlier patterns effectively. Additional inference results can be seen in Figure A3 in Appendix B.

Figure 15. (**a**–**h**) Example outputs of the outlier object detection model. The detected outliers are marked with red bounding boxes, with the confidence scores displayed above each box.
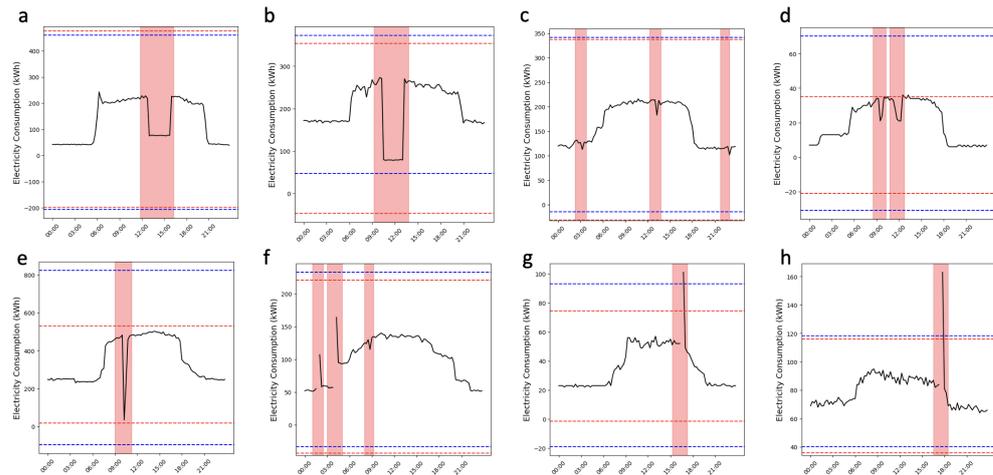
5. Comparison and Discussion

To facilitate the comparison between the outcomes from the VOD method and those of the baseline outlier detection methods, a procedure for transforming the results of object detection bounding boxes into corresponding time periods is proposed. As illustrated in Figure 16, the process begins with the extraction of bounding boxes from the object detection model. The X-axis pixel coordinates of these boxes are then mapped onto their respective timestamps within a daily timeframe. This mapping is based on the correlation between the pixel coordinates and timestamps on the X-axis. Finally, the temporal intervals are highlighted to clearly demonstrate the periods identified with outlier occurrences.
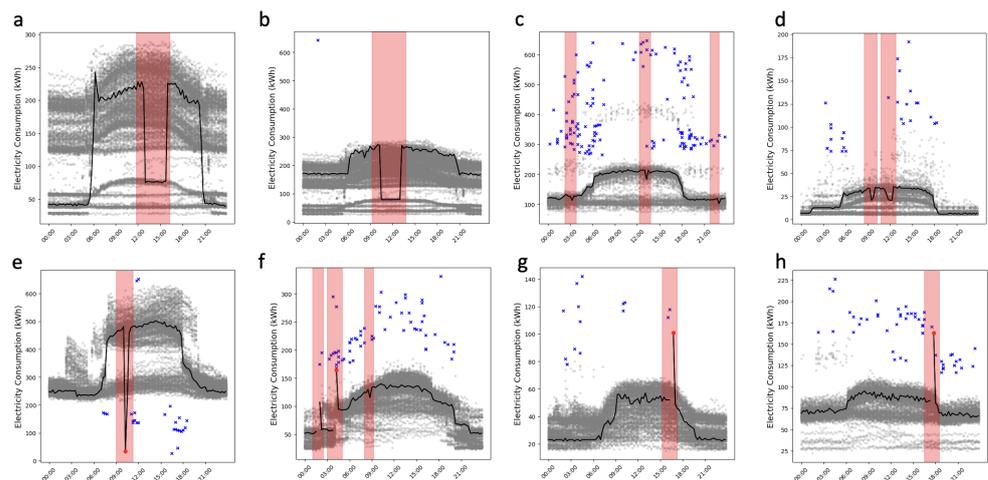

Figure 16. The process of transforming bounding boxes into corresponding temporal intervals.

Two baseline methods, IQR and DBSCAN, are used to benchmark the performance of the VOD method. The analysis utilizes data on electricity consumption over eight workdays to evaluate the outlier detection performance across different scenarios. Figure 17 illustrates the comparative results between the VOD and IQR methods. For the IQR method, in scenarios a through f, neither the daily nor the annual boundaries could effectively detect outliers. However, both boundaries accurately identify the point outlier in scenarios g and h. This suggests that the IQR method is good at detecting the "extreme" point outlier when the data record significantly deviates far from the norm, while not good at detecting the contextual outliers when the aberration is not in the data scale but in the timing of the values. Conversely, the VOD method demonstrates its efficacy by identifying collective outliers in scenarios a, b, and d; contextual outliers in scenarios c and f; and point outliers in scenarios e, f, g, and h.

Figure 17. (**a**–**h**) Comparison between VOD and IQR methods. The black lines are the daily building electricity energy consumption time-series plots. The areas highlighted in red indicate the temporal intervals flagged by the VOD method due to the presence of outliers. The red and blue lines are the "1.5 IQR" boundaries derived from the annual and daily electricity consumption data, respectively.

Figure 18 illustrates the comparative results between the VOD and DBSCAN methods. To contextualize the DBSCAN detection outcomes, the entire year's electricity consumption data are provided in the background, with outliers identified by the DBSCAN method marked by "x" symbols. The DBSCAN method slightly improves over the IQR method by successfully identifying outliers in scenarios e and f. Similar to the IQR method, the DBSCAN method underperforms in scenarios a to d. Different from the simple horizontal thresholds from the IQR method, DBSCAN ascertains outliers through the spatial proximity between each data point. However, its limitation becomes apparent in situations where ample historical data validate the occurrence of outliers—when there is a significant amount of data from a similar timeframe with similar value scales. The method also treats data points independently, regardless of whether they occur on the same day, and overlooks the correlation among daily data points. Consequently, the method cannot fully capture the building energy consumption "load shape", thereby compromising its effectiveness in scenarios a to d.

Figure 18. (**a**–**h**) Comparison between VOD and DBSCAN methods. The black lines are the daily building electricity energy consumption time-series plots for the daily testing scenarios. The areas highlighted in red indicate the temporal intervals flagged by the VOD method due to the presence of outliers. In the background, gray dots represent the plots of aggregated annual electricity consumption data on a daily basis. Outliers detected by the DBSCAN method from the annual dataset are marked with blue "x" symbols, while those in the tested daily scenarios are specifically highlighted with red dots.

6. Conclusions

This study has successfully developed and demonstrated a vision-based outlier detection (VOD) method for detecting outliers in building energy data by innovatively transforming time-series energy data into 2D time-series plots. The idea is inspired by the observation that humans can easily point out outliers by looking at time-series plots. The expertise in detecting building energy outliers is captured in the labeled datasets, and the deep learning models are trained with these datasets, effectively transferring the expert knowledge to the model.

The VOD method is an experiment in learning and emulating human expertise in pattern and outlier recognition within time-series data plots. This method has significant potential for application in other time-series data within the building area, such as Indoor Air Quality (IAQ) and Internet of Things (IoT) data, offering a versatile tool for diverse building applications. However, the success of this technique is heavily dependent on the quality and quantity of the labeled data used for training. Therefore, enhancing the process of data collection and labeling is crucial for the continual improvement of the model's accuracy and reliability.

Further research directions include a deeper exploration into the underlying causes of different outliers, aiming to categorize them into more distinct groups. While the current study focuses primarily on daily energy consumption data during workdays, forthcoming research could broaden its scope to encompass data from all days and evenings and extend the detection window from a daily scale to weekly, monthly, and even yearly intervals. This expansion would not only refine the outlier detection capability but also yield more comprehensive insights into building energy usage patterns, aiding in the development of more efficient and sustainable energy management strategies in buildings.

Appendix A. Grad-CAM Visualization of the Classification Model

Appendix A.1. Visualization of the "Looking Good" Category

Figure A1. Visualization of important regions for the "looking good" category. The red regions in the heatmaps indicate the most informative parts used by the model for the prediction of "looking good".

Appendix A.2. Visualization of the " Potential Problems" Category

Figure A2. Visualization of important regions for the "potential problems" category. The red regions in the heatmaps indicate the most informative parts used by the model for the prediction of "potential problems".

Appendix B. Validation and Test Results of the Outlier Object Detection Model
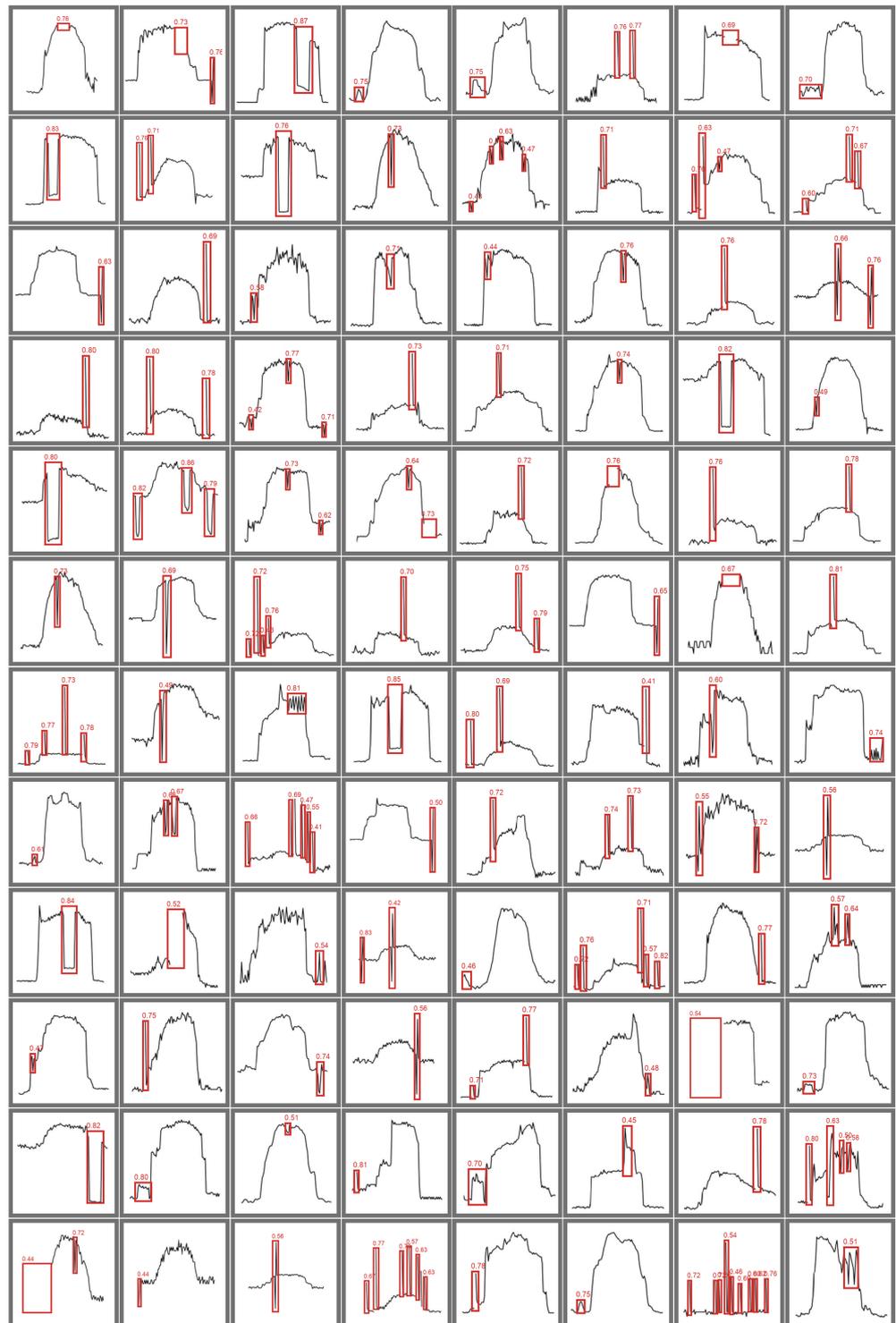

Figure A3. Validation and test results of the outlier object detection model. Confidence score limit: 0.4; IoU threshold: 0.3.

References

1. EIA. *Buildings Sectorial Overview*; Technical Report CC BY 4.0; U.S. Energy Information Administration (EIA): Washington, DC, USA, 2022.
2. Li, T.; Liu, T.; Sawyer, A.O.; Tang, P.; Loftness, V.; Lu, Y.; Xie, J. Generalized building energy and carbon emissions benchmarking with post-prediction analysis. *Dev. Built Environ.* **2024**, *17*, 100320. [CrossRef]
3. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–58. [CrossRef]
4. Gaddam, A.; Wilkin, T.; Angelova, M.; Gaddam, J. Detecting sensor faults, anomalies and outliers in the internet of things: A survey on the challenges and solutions. *Electronics* **2020**, *9*, 511. [CrossRef]
5. Xu, C.; Zhao, S.; Liu, F. Sensor fault detection and diagnosis in the presence of outliers. *Neurocomputing* **2019**, *349*, 156–163. [CrossRef]
6. Li, X.; Bowers, C.P.; Schnier, T. Classification of energy consumption in buildings with outlier detection. *IEEE Trans. Ind. Electron.* **2009**, *57*, 3639–3644. [CrossRef]
7. Martin Nascimento, G.F.; Wurtz, F.; Kuo-Peng, P.; Delinchant, B.; Jhoe Batistela, N. Outlier Detection in Buildings' Power Consumption Data Using Forecast Error. *Energies* **2021**, *14*, 8325. [CrossRef]
8. Larson, S.; Mahendran, A.; Lee, A.; Kummerfeld, J.K.; Hill, P.; Laurenzano, M.A.; Hauswald, J.; Tang, L.; Mars, J. Outlier detection for improved data quality and diversity in dialog systems. *arXiv* **2019**, arXiv:1904.03122. [CrossRef]
9. Zhang, Y.; Meratnia, N.; Havinga, P.J. Ensuring high sensor data quality through use of online outlier detection techniques. *Int. J. Sens. Netw.* **2010**, *7*, 141–151. [CrossRef]
10. Himeur, Y.; Ghanem, K.; Alsalemi, A.; Bensaali, F.; Amira, A. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Appl. Energy* **2021**, *287*, 116601. [CrossRef]
11. Copiaco, A.; Himeur, Y.; Amira, A.; Mansoor, W.; Fadli, F.; Atalla, S.; Sohail, S.S. An innovative deep anomaly detection of building energy consumption using energy time-series images. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105775. [CrossRef]
12. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd* **1996**, *96*, 226–231.
13. Jalori, S.; Reddy, T.A. A New Clustering Method to Identify Outliers and Diurnal Schedules from Building Energy Interval Data. *ASHRAE Trans.* **2015**, *121*, 33.
14. Piscitelli, M.S.; Brandi, S.; Capozzoli, A. Recognition and classification of typical load profiles in buildings with non-intrusive learning approach. *Appl. Energy* **2019**, *255*, 113727. [CrossRef]
15. Liu, X.; Sun, H.; Han, S.; Han, S.; Niu, S.; Qin, W.; Sun, P.; Song, D. A data mining research on office building energy pattern based on time-series energy consumption data. *Energy Build.* **2022**, *259*, 111888. [CrossRef]
16. Li, T.; Bie, H.; Lu, Y.; Sawyer, A.O.; Loftness, V. MEBA: AI-powered precise building monthly energy benchmarking approach. *Appl. Energy* **2024**, *359*, 122716. [CrossRef]
17. Price, P. *Methods for Analyzing Electric Load Shape and Its Variability*; Technical Report; Lawrence Berkeley National Lab. (LBNL): Berkeley, CA, USA, 2010. [CrossRef]
18. Luo, X.; Hong, T.; Chen, Y.; Piette, M.A. Electric load shape benchmarking for small-and medium-sized commercial buildings. *Appl. Energy* **2017**, *204*, 715–725. [CrossRef]
19. Liu, X.; Ding, Y.; Tang, H.; Xiao, F. A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. *Energy Build.* **2021**, *231*, 110601. [CrossRef]
20. Fan, C.; Xiao, F.; Zhao, Y.; Wang, J. Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data. *Appl. Energy* **2018**, *211*, 1123–1135. [CrossRef]
21. Zheng, Z.; Yang, Y.; Niu, X.; Dai, H.N.; Zhou, Y. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Trans. Ind. Inform.* **2017**, *14*, 1606–1615. [CrossRef]
22. Fahim, M.; Fraz, K.; Sillitti, A. TSI: Time series to imaging based model for detecting anomalous energy consumption in smart buildings. *Inf. Sci.* **2020**, *523*, 1–13. [CrossRef]
23. Carcillo, F.; Le Borgne, Y.A.; Caelen, O.; Kessaci, Y.; Oblé, F.; Bontempi, G. Combining unsupervised and supervised learning in credit card fraud detection. *Inf. Sci.* **2021**, *557*, 317–331. [CrossRef]
24. Bawono, A.H.; Bachtiar, F.A. Outlier Detection with Supervised Learning Method. In Proceedings of the 2019 International Conference on Sustainable Information Engineering and Technology (SIET), Lombok, Indonesia, 28–30 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 306–309. [CrossRef]
25. Paulheim, H.; Meusel, R. A decomposition of the outlier detection problem into a set of supervised learning problems. *Mach. Learn.* **2015**, *100*, 509–531. [CrossRef]
26. Aggarwal, C.C.; Aggarwal, C.C. Supervised outlier detection. In *Outlier Analysis*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 219–248.
27. Takahashi, K.; Ooka, R.; Kurosaki, A. Seasonal threshold to reduce false positives for prediction-based outlier detection in building energy data. *J. Build. Eng.* **2024**, *84*, 108539. [CrossRef]
28. Araya, D.B.; Grolinger, K.; ElYamany, H.F.; Capretz, M.A.; Bitsuamlak, G. An ensemble learning framework for anomaly detection in building energy consumption. *Energy Build.* **2017**, *144*, 191–206. [CrossRef]

29. Shoemaker, L.; Hall, L.O. Anomaly detection using ensembles. In Proceedings of the Multiple Classifier Systems: 10th International Workshop, MCS 2011, Naples, Italy, 15–17 June 2011; Proceedings 10; Springer: Berlin/Heidelberg, Germany, 2011; pp. 6–15. [CrossRef]

30. Miyata, S.; Akashi, Y.; Lim, J.; Motomura, A.; Tanaka, K.; Tanaka, S.; Kuwahara, Y. Fault Detection and Diagnosis in Building Heat Source Systems Using Machine Learning (Part 2) Preprocessing of Fault Data for Improvement in Diagnosis Performance and Application to BEMS Data. *Trans. SHASE Jpn.* **2018**, *261*, 1–9.

31. Xu, C.; Chen, H. Abnormal energy consumption detection for GSHP system based on ensemble deep learning and statistical modeling method. *Int. J. Refrig.* **2020**, *114*, 106–117. [CrossRef]

32. Hsu, D. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. *Appl. Energy* **2015**, *160*, 153–163. [CrossRef]

33. Chen, Z.; Xiao, F.; Guo, F.; Yan, J. Interpretable machine learning for building energy management: A state-of-the-art review. *Adv. Appl. Energy* **2023**, *9*, 100123. [CrossRef]

34. Choo, J.; Liu, S. Visual analytics for explainable deep learning. *IEEE Comput. Graph. Appl.* **2018**, *38*, 84–92. [CrossRef]

35. Li, Y.; O'Neill, Z.; Zhang, L.; Chen, J.; Im, P.; DeGraw, J. Grey-box modeling and application for building energy simulations—A critical review. *Renew. Sustain. Energy Rev.* **2021**, *146*, 111174. [CrossRef]

36. Wang, Z.; Hong, T.; Piette, M.A. Building thermal load prediction through shallow machine learning and deep learning. *Appl. Energy* **2020**, *263*, 114683. [CrossRef]

37. Machlev, R.; Heistrene, L.; Perl, M.; Levy, K.; Belikov, J.; Mannor, S.; Levron, Y. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy AI* **2022**, *9*, 100169. [CrossRef]

38. Somu, N.; MR, G.R.; Ramamritham, K. A deep learning framework for building energy consumption forecast. *Renew. Sustain. Energy Rev.* **2021**, *137*, 110591. [CrossRef]

39. Li, C.; Ding, Z.; Zhao, D.; Yi, J.; Zhang, G. Building energy consumption prediction: An extreme deep learning approach. *Energies* **2017**, *10*, 1525. [CrossRef]

40. Fan, C.; Xiao, F.; Zhao, Y. A short-term building cooling load prediction method using deep learning algorithms. *Appl. Energy* **2017**, *195*, 222–233. [CrossRef]

41. Hawkins, D.M. *Identification of Outliers*; Springer: Berlin/Heidelberg, Germany, 1980; Volume 11. [CrossRef]

42. Cook, A.A.; Mısırlı, G.; Fan, Z. Anomaly detection for IoT time-series data: A survey. *IEEE Internet Things J.* **2019**, *7*, 6481–6494. [CrossRef]

43. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef] [PubMed]

44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

45. Shafiq, M.; Gu, Z. Deep residual learning for image recognition: A survey. *Appl. Sci.* **2022**, *12*, 8972. [CrossRef]

46. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

47. Jocher, G.; Stoken, A.; Borovec, J.; Changyu, L.; Hogan, A.; Diaconu, L.; Poznanski, J.; Yu, L.; Rai, P.; Ferriday, R.; et al. ultralytics/yolov5: v3. 0. *Zenodo* **2020**. [CrossRef]

48. Wang, Z.; Jin, L.; Wang, S.; Xu, H. Apple stem/calyx real-time recognition using YOLO-v5 algorithm for fruit automatic loading system. *Postharvest Biol. Technol.* **2022**, *185*, 111808. [CrossRef]

49. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.

50. Zhou, F.; Zhao, H.; Nie, Z. Safety helmet detection based on YOLOv5. In Proceedings of the 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), Shenyang, China, 22–24 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 6–11. [CrossRef]

51. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768. [CrossRef]

52. Overview of Model Structure about YOLOv5. Available online: https://github.com/ultralytics/yolov5/issues/280 (accessed on 30 March 2024).

53. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755. [CrossRef]

54. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]