sustainability

MDPI

Article

DSRA-DETR: An Improved DETR for Multiscale Traffic Sign Detection

Jiaao Xia, Meijuan Li, Weikang Liu and Xuebo Chen *

School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan 114051, China
* Correspondence: xuebochen@126.com

Abstract: Traffic sign detection plays an important role in improving the capabilities of automated driving systems by addressing road safety challenges in sustainable urban living. In this paper, we present DSRA-DETR, a novel approach focused on improving multiscale detection performance. Our approach integrates the dilated spatial pyramid pooling model (DSPP) and the multiscale feature residual aggregation module (FRAM) to aggregate features at various scales. These modules excel at reducing feature noise and minimizing loss of low-level features during feature map extraction. Additionally, they enhance the model's capability to detect objects at different scales, thereby improving the accuracy and robustness of traffic sign detection. We evaluate the performance of our method on two widely used datasets, the GTSDB and CCTSDB, and achieve impressive average accuracies (APs) of 76.13% and 78.24%, respectively. Compared with other well-known algorithms, our method shows a significant improvement in detection accuracy, demonstrating its superiority and generality. Our proposed method shows great potential for improving the performance of traffic sign detection for autonomous driving systems and will help in the development of safe and efficient autonomous driving technologies.

Keywords: traffic sign detection; autonomous driving systems; pyramid pooling; DSRA-DETR

Citation: Xia, J.; Li, M.; Liu, W.; Chen, X. DSRA-DETR: An Improved DETR for Multiscale Traffic Sign Detection. *Sustainability* **2023**, *15*, 10862. https://doi.org/10.3390/su151410862

Academic Editors: Haoran Wei, Zhendong Wang, Yuchao Chang and Zhenghua Huang

Received: 22 May 2023
Revised: 6 July 2023
Accepted: 10 July 2023
Published: 11 July 2023

1. Introduction

In sustainable urban living, road safety faces challenges such as distracted drivers and novice drivers' unfamiliarity with traffic signs. To address these challenges, traffic sign detection technology in autonomous driving systems can assist drivers in identifying traffic signs accurately, contributing to road safety. Achieving accurate recognition of small-sized traffic signs is crucial for autonomous vehicles to assess road conditions and ensure safe operation. Researchers are actively working to improve the detection accuracy of small-scale traffic sign images, aiming to enhance the performance and reliability of autonomous driving systems.

Accurate recognition of small-scale traffic signs is essential for advancing autonomous driving technology, providing autonomous vehicles with sufficient time to respond to changing road conditions. Detecting and interpreting small-scale traffic signs accurately contribute to the safety, efficiency, and reliability of autonomous vehicles, making it a key research area within AI applications for sustainable urban living [1]. Deep learning methods, including the R-CNN [2] series, YOLO [3] series, SSD [4] series, and visual transformer architecture [5], have been widely used for traffic sign detection. The introduction of DETR [6] has paved the way for transformer-based target detectors. However, the existing methods still have limitations in terms of detection accuracy and the detection of small traffic signs at a distance.

In this paper, we propose a novel traffic sign detection method called DSRA-DETR, which improves upon Anchor-DETR [7] by integrating designed modules. DSRA-DETR utilizes multiscale feature information extracted from the backbone and employs dilated

spatial pyramid pooling (DSPP) and the feature residual aggregation module (FRAM) to enhance the feature map's representational power. Experimental results for the GTSDB [8] and CCTSDB [9] datasets demonstrate that DSRA-DETR outperforms existing methods in terms of accuracy.

Our contributions include applying Anchor-DETR to traffic sign detection, introducing the FRAM for multiscale feature aggregation, and proposing the DSPP module for enhanced feature representation. The proposed DSRA-DETR method achieves better accuracy compared to YOLOv3 [10] and Conditional DETR [6]. The remaining sections of this paper provide an overview of the related literature, a detailed description of the methodology, information about the datasets and evaluation metrics, experimental setup and results, and a summary of findings and future research directions.

2. Related Work

2.1. The R-CNN Series

The R-CNN series is considered a two-stage algorithm, with R-CNN [2] being a pioneering attempt to apply convolutional neural networks (CNNs) for target detection. It leverages a CNN to extract region proposals in features and subsequently employs SVM classification with bbox regression. To address the slow speed issue of R-CNN, He et al. proposed SPPNet [11], which incorporates an SPP layer between the final convolutional layer and the fully connected layer. Building upon the concepts in SPPNet, Girshick et al. enhanced R-CNN and introduced Fast R-CNN [12], unifying category judgment and position regression through a deep network implementation and thereby improving testing and training speed. An upgraded version of Fast R-CNN is Faster R-CNN [13], which integrates the four fundamental steps of target detection (candidate region generation, feature extraction, classification, and location refinement) into a comprehensive deep network framework. Instead of using the original SS (region proposals), it employs an RPN (region proposal network). Additionally, following the introduction of ResNet [14], He et al. combined parts of the Faster R-CNN architecture, introduced RoI Align to replace RoI pooling, and proposed Mask R-CNN [15], which yielded superior performance and greater scalability. Some scholars [16,17] use this series of methods for traffic sign detection.

2.2. The YOLO and SSD Series

The YOLO series is considered a standard one-stage algorithm. YOLOv1 [3], which is the first paper in this series, introduced the core idea of using the entire image as the network input and directly regressing the location and category of bounding boxes in the output layer. However, it falls short in terms of localization accuracy compared to Faster R-CNN and struggles with detecting small objects. YOLOv2 [18], an advancement over the v1 version, addresses these limitations in three key aspects: improved prediction accuracy, faster processing speed, and enhanced object recognition, all while maintaining its efficient processing speed. YOLOv3 [10] further enhances the architecture and training techniques introduced in v2 to improve accuracy without compromising inference time. In addition, refs. [19,20] have made notable contributions in further enhancing the YOLO algorithm. Liu et al. introduced the SSD [4] algorithm, which is based on multiscale detection. It achieves a processing speed comparable to that of YOLO and a detection accuracy comparable to that of Faster R-CNN. However, its performance in detecting small targets is still not entirely satisfactory. Both [21–23] have made improvements to the SSD algorithm from different angles. It is worth mentioning that RetinaNet [24] proposes focal loss to address the issue of severe imbalance between positive and negative sample ratios in one-stage target detection. Moreover, scholars have introduced the CornerNet [25] algorithm, which utilizes diagonal keypoints to tackle the bounding box(bbox) problem. Building upon this, CenterNet [26], employs central keypoints to further address the bbox problem. Some scholars [27–29] use this series of methods for traffic sign detection.

2.3. The Image Registration Series

In traffic sign detection, feature descriptors in image alignment can be used to extract features in traffic sign images and match them with corresponding features in other images. Among these, it is worth mentioning the FNRG [30] method proposed by Xiao et al., which starts with a novel consistency seed search strategy. This strategy exploits the first neighbor relationship of feature points between two images to achieve consistency matching without any parameters or thresholds. It is an eye-catching image-matching method. Additionally, the LGF algorithm consists of two components: an effective two-view approximate deterministic sampling algorithm and a simple and effective model selection framework. The LGF [31] algorithm is able to obtain a coarse minimum subset of samples using the local neighbor-keeping relationships corresponding to the inputs. It then refines these subsets using a global residual optimization strategy. In this way, the same traffic signs appearing in different images can be detected. Some scholars [32,33] use this series of methods for traffic sign detection.

2.4. The DETR Series

More recently, the detection transformer (DETR) [34] became the first architecture to apply the transformer [35] architecture to target detection, marking a significant advancement in the vision field. While it demonstrates impressive performance on the COCO [36] dataset, its convergence speed is relatively slow due to the computational demands of the transformer architecture [5]. To tackle this issue, Deformable-DETR [37] proposes a deformable attention mechanism and Conditional DETR [6] introduces a conditional cross-attention mechanism, both of which make important contributions in reducing the convergence time of DETR. Many scholars, including [38–40], have made significant contributions to enhancing DETR by introducing improvements from various perspectives and degrees based on the aforementioned work. Anchor-DETR [7], on the other hand, is a deformable-based object detection framework that incorporates anchor points and row–column decouple attention into DETR. Anchor-DETR is known for its fast convergence and competitive performance compared to other detectors.

3. Method

DSRA-DETR is an advanced traffic sign detection architecture that builds upon the foundation of Anchor-DETR. In order to overcome the challenges posed by small-scale traffic sign detection, DSRA-DETR introduces a series of innovative components. One such component is the dilated spatial pyramid pooling module, which plays a crucial role in this architecture. By leveraging dilated convolutions, this module effectively filters out extraneous and irrelevant information from the low-level features. This filtering process ensures that only the most relevant and discriminative features are retained for further analysis and processing. Additionally, DSRA-DETR incorporates a feature residual aggregation module, which serves as a vital component for aggregating and enhancing the representation of low-level feature information. This module intelligently combines and refines the extracted features, enabling the model to capture more detailed and context-aware representations of traffic signs. By integrating this module into the architecture, DSRA-DETR significantly improves the accuracy and robustness of traffic sign detection, particularly in scenarios where small-scale signs are prevalent.

Figure 1 provides a visual overview of the DSRA-DETR architecture, showcasing its various components and their interactions. The backbone network forms the foundation of the architecture, being responsible for extracting initial feature representations from the input data. The dilated spatial pyramid pooling model operates on these features, capturing multiscale information and selectively incorporating contextual details. The feature residual aggregation module then refines the features, enhancing their discriminative power and contributing to the overall performance of the model. After being processed by this module, the feature is then fed into the encoder layer and decoder layer of the transformer.

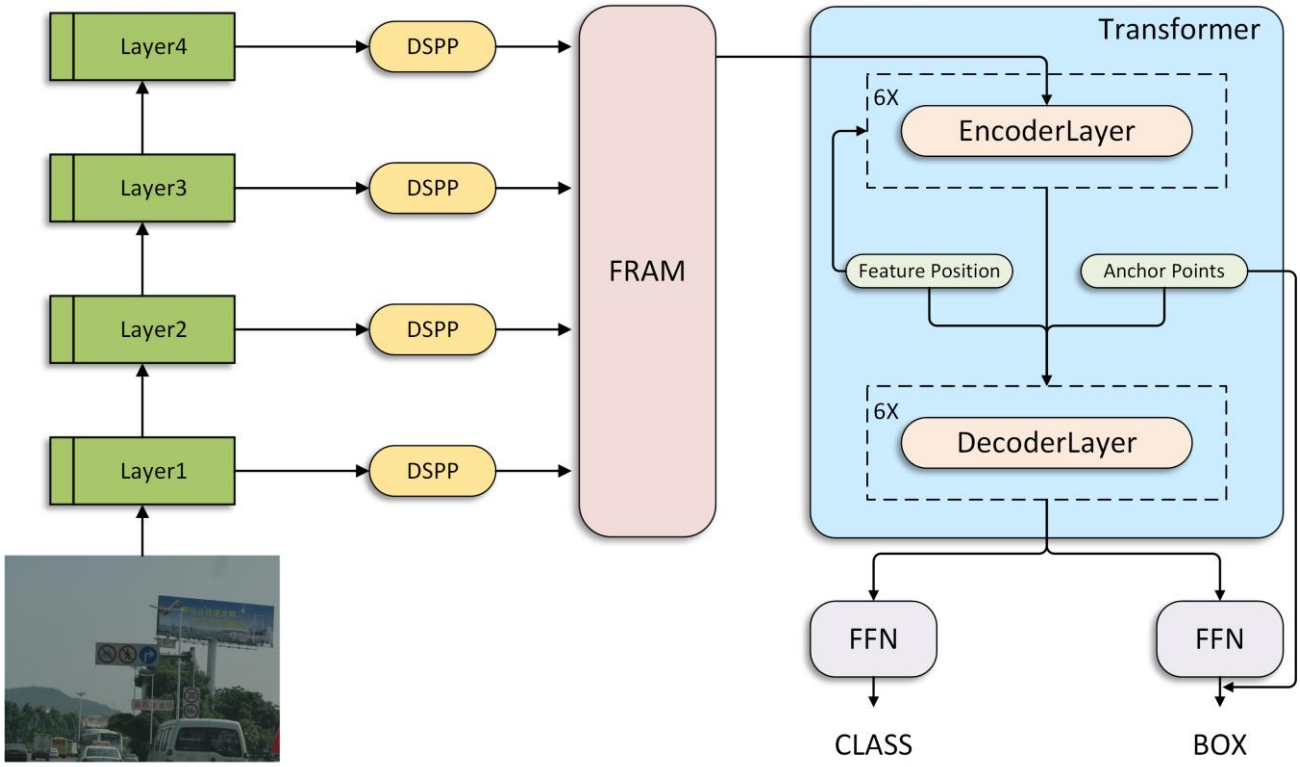

Figure 1. Dilated spatial residual Anchor-DETR.

3.1. Backbone

The backbone plays a crucial role in the target detection task by aiding the model in extracting features from the input image. These features are then utilized in the latter part of the model. As a result, having a strong backbone is essential for our traffic sign detection task. We use ResNet50 as the backbone network for all models, which are pre-trained on ImageNet. Figure 2 illustrates the structure of the backbone we are using.

ResNet50 consists of four layers: layer1, layer2, layer3, and layer4. Each layer follows the same internal structure but has different downsampling rates: 4×, 8×, 16×, and 32×, respectively. The lower layers contain detailed location feature information, making them suitable for detecting small targets. On the other hand, the higher layers capture abstract semantic features, making them more suitable for detecting larger targets. In our method, we leverage the multilayer features extracted from all four layers for further processing and analysis.

3.2. Dilated Spatial Pyramid Pooling Model

The DSPP module is an essential element of our proposed model, which draws inspiration from the design principles of the DeepLabv2 [41] architecture. However, we made several improvements to make it more suitable for our specific needs. The module comprises four convolutional layers: three 3×3 dilated convolutional layers with expansion rates of [1,3,6] and one 1×1 convolutional layer. The use of dilated convolutional layers instead of regular convolutional layers reduces the computational cost of the module while maintaining its effectiveness.

To apply the DSPP module to the input features, we first convolve the input with three different expansion rates in parallel. The expansion rate refers to the number of output channels per input channel. Then, we concatenate the resulting feature maps before passing them through the final 1×1 convolutional layer, which downscales the feature maps to a desired number of output channels. The DSPP module allows our model to capture features at multiple scales, allowing for more accurate detection of traffic signs of varying sizes and scales. The module's structure is depicted in Figure 3.

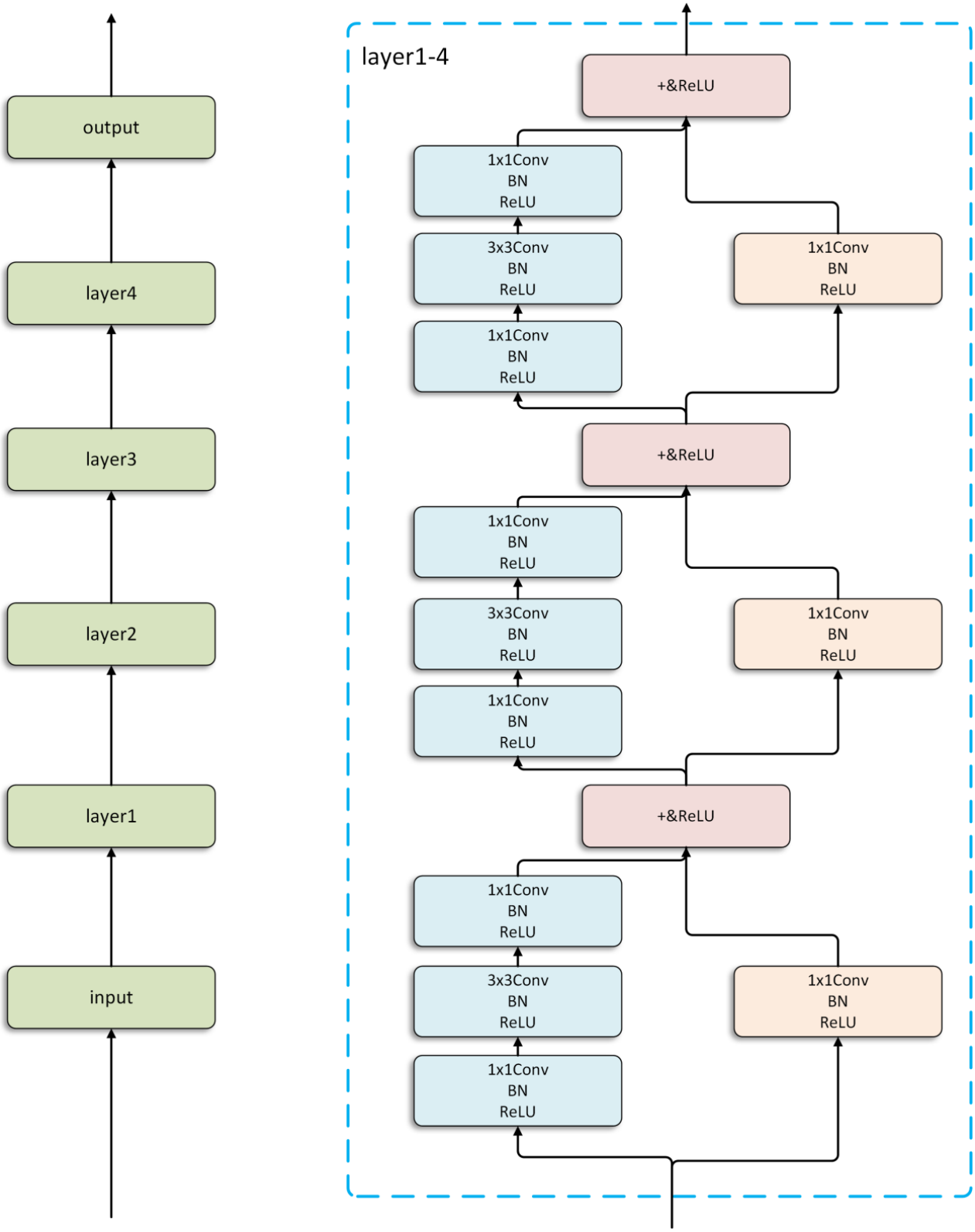

Figure 2. The structure of ResNet50.

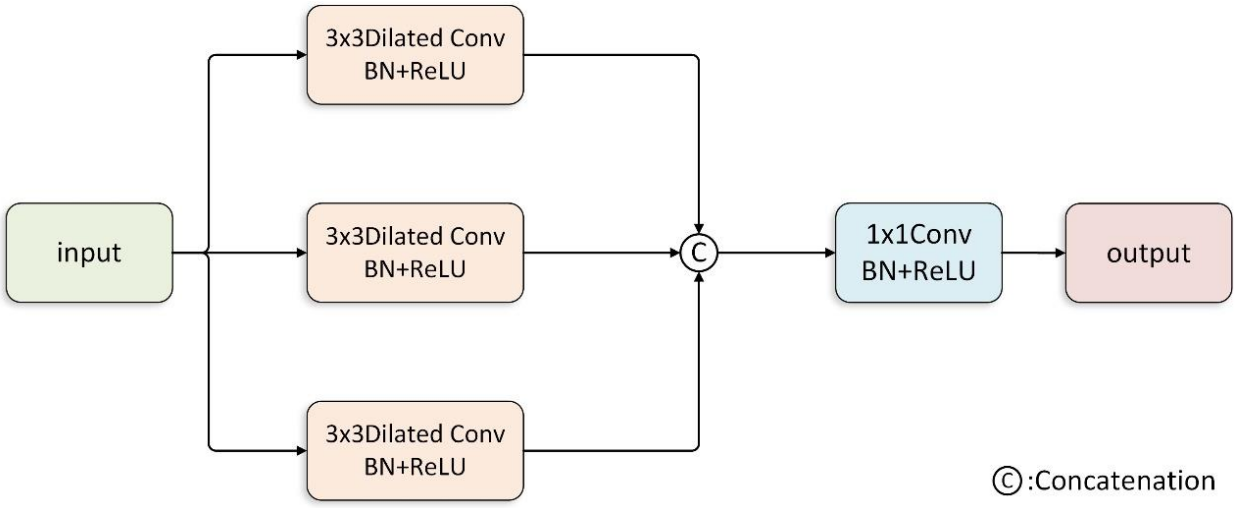

Figure 3. Dilated spatial pyramid pooling module.

The input feature map, F_{in}, has a shape of $F_{in} \in R^{B\times C\times H\times W}$, where B represents batches, C represents channels, and H and W represent the height and width, respectively. The mathematical expression for F_{in} after passing through a dilated convolution block with a dilation rate of i is as follows:

$$F_i = ReLU(BN(Conv_{3\times 3}(F_{in}))) \tag{1}$$

In this expression, $ReLU$ refers to the $ReLU$ activation function, BN represents batch normalization, and $Conv_{3\times 3}$ denotes a 3×3 dilated convolution operation. The overall expression for this module can be expressed as follows:

$$F_{out} = ReLU(BN(Conv_{1\times 1}(concat(F_1, F_3, F_6)))) \tag{2}$$

Here, F_{out} represents the resulting output feature map. It is obtained by concatenating feature maps F_1, F_3, and F_6, and then applying a 1×1 convolution, followed by batch normalization and $ReLU$ activation.

Through the utilization of the DSPP module, we are able to apply targeted noise reduction to the feature information extracted from the backbone. This noise reduction process selectively preserves the relevant traffic sign features while removing unwanted "feature noise" from the feature map. Consequently, the original feature map from the backbone can more effectively carry out the task of traffic sign detection following the integration of the DSPP module. In our ablation experiments, we visually demonstrated the impact of this process by visualizing the feature maps, providing a clearer and more illustrative understanding of this point.

3.3. Feature Residual Aggregation Module

The feature residual aggregation module (FRAM) is a crucial component in our proposed model architecture. It addresses the challenge of feature resolution discrepancies encountered in object detection tasks. The FRAM effectively preserves and leverages lower-level features, resulting in significant improvements in the model's detection performance, especially for small-scale traffic signs. Its primary objective is to ensure that the model retains essential information from lower-level features while extracting them hierarchically. This is achieved through a feature residual aggregation process that considers features from different scale layers that have undergone the DSPP module.

Inside the FRAM, the process starts with an absolute value subtraction of the feature matrix at each level. This step calculates the differences between the feature layers from different levels, allowing the module to discern disparities in content and characteris-

tics. The differences obtained from the layer-wise calculations are then convolved with the original high-level features. This convolution operation integrates the dissimilarities between the layers with the existing high-level features, resulting in a comprehensive representation of the combined information. By fusing the residuals of the low-level features, richer information on small target features is aggregated in the feature maps used in the subsequent detection part. This is crucial for the improvement in small-target detection performance. To ensure successful fusion, the convolved results from different layers are concatenated. This concatenation step consolidates the information obtained from each layer and prepares it for subsequent processing. The concatenated features undergo a downsampling operation, reducing the number of output channels to the desired level.

By utilizing the FRAM, our model demonstrates notable improvements in detection capabilities, particularly in recognizing small-scale traffic signs. The module preserves and effectively leverages essential information from lower-level features, enabling the model to capture and utilize intricate details associated with traffic signs more efficiently. Handling feature resolution discrepancies and preserving critical information from lower-level features enhances the model's accuracy and robustness. This enhancement enables the model to detect traffic signs of varying sizes and scales with greater precision and reliability. Please refer to Figure 4 for the structure diagram.

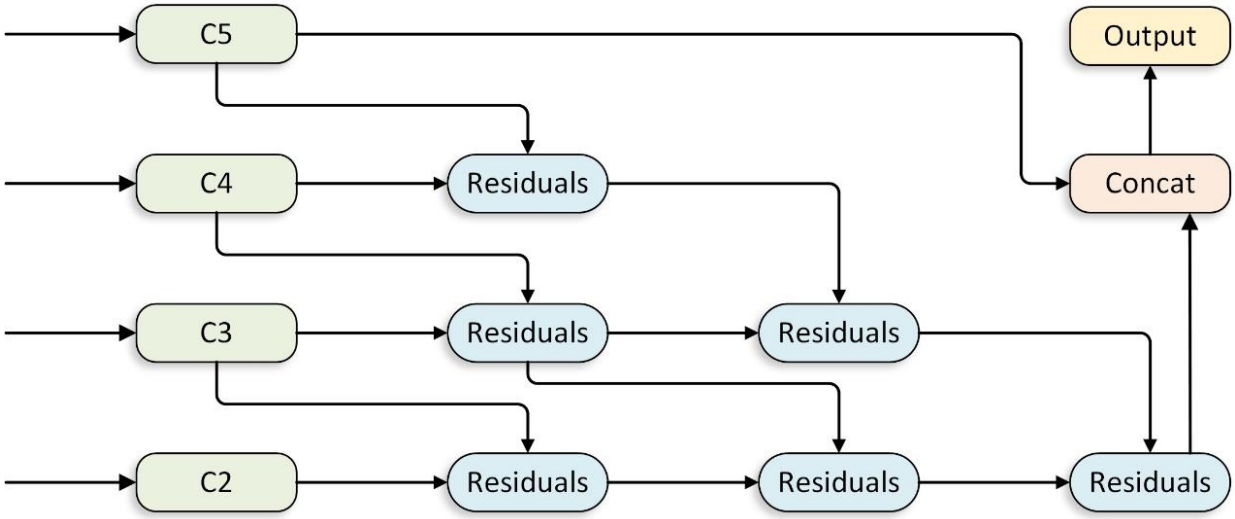

Figure 4. Feature residual aggregation module.

In this module, $C2$, $C3$, $C4$, and $C5$ represent the feature maps of different layers after passing through the DSPP module. Their shapes are as follows: $C2 \in R^{B\times C\times \frac{H}{4}\times \frac{W}{4}}$, $C3 \in R^{B\times C\times \frac{H}{8}\times \frac{W}{8}}$, $C4 \in R^{B\times C\times \frac{H}{16}\times \frac{W}{16}}$, $C5 \in R^{B\times C\times \frac{H}{32}\times \frac{W}{32}}$. The calculation formula for the residuals module can be expressed as

$$F_r = Abs(Ds(F_h) - F_l) \tag{3}$$

Here, F_r represents the output of a residual module. F_h denotes the high-level feature input received by the module, F_l denotes the low-level feature input received by the module, Ds represents the downsampling operation, and Abs represents the absolute value operation. Therefore, the overall calculation formula for this module can be expressed as

$$F_{out} = ReLU(BN(Conv_{1\times 1}(concat(C5, R_{5,4,3,2})))) \tag{4}$$

In the formula, $R_{5,4,3,2}$ represents the residuals obtained by aggregating the respective layer features.

3.4. Losses

The loss function has a significant impact on target detection. Its purpose is to assess the disparity or error between the predicted outcomes of the model and the real labels. By quantifying the difference between the predicted value and the true label, the loss function offers feedback signals to guide the model's optimization and learning during training. To accurately detect objects in an image, we utilize a combination of classification and bounding box regression tasks. The classification task involves predicting the labels of the objects present in the image, while the bounding box regression task aims to accurately locate the objects in the image. To supervise these tasks, we use appropriate loss functions. Specifically, for the classification task, we use the cross-entropy loss function, which is given by

$$L_{class} = -ylog(p) - (1-y)log(1-p) \tag{5}$$

Here, y denotes the true label of the sample and p denotes the predicted probability of the model. For the bounding box loss, we use a linear combination of L1 loss and Generalized Intersection over Union (GIoU) Loss:

$$L_{box}(X,Y) = \lambda_{iou}L_{iou}(X,Y) + \lambda_{L1}(X,Y) \tag{6}$$

where $L_{iou}(X,Y)$ denotes the GIoU loss function, $\lambda_{L1}(X,Y)$ represents the L1 distance loss function, and λ_{iou} and λ_{L1} are hyperparameters. The L1 loss function is defined as

$$L_1(X,Y) = \mid X - Y \mid \tag{7}$$

Furthermore, the GIoU loss function is given by

$$L_{iou} = 1 - \left(\frac{\mid X \cap Y \mid}{\mid X \cup Y \mid} - \frac{\mid C \setminus (X \cup Y) \mid}{\mid C \mid} \right) \tag{8}$$

Here, X and Y denote the true and predicted bounding boxes, respectively, and C represents the minimum bounding box containing X and Y. The symbols $|.|$ and $\setminus$ indicate the area and set differences, respectively. Finally, our overall loss function is expressed as

$$Loss = \sum_{i=1}^{N} [L_{class} + L_{box}(X,Y)] \tag{9}$$

where N is the number of samples in the batch. This loss function combines the classification and bounding box losses, encouraging the model to make accurate predictions for both tasks simultaneously. By minimizing this loss function during training, the model learns to accurately classify and locate objects in images.

4. Dataset

4.1. Datasets

This study utilizes two well-established datasets, namely the German Traffic Sign Detection Dataset (GTSDB) and the Chinese Traffic Sign Dataset (CCTSDB), to facilitate the evaluation of traffic sign detection algorithms. The GTSDB consists of a total of 827 images, each with a resolution of 800 $\times$ 1360 pixels, encompassing four distinct types of traffic signs: "prohibitory", "mandatory", "danger", and "other". The size of the traffic signs in this dataset varies from 16 to 126 pixels. On the other hand, the CCTSDB comprises 17,856 images, with resolutions of 760 $\times$ 1280 and 768 $\times$ 1024 pixels, and includes three types of traffic signs: "prohibitory", "warning", and "mandatory". These datasets provide a comprehensive collection of traffic sign images captured in Germany and China, offering diverse variations in weather conditions and road scenarios. To provide visual exemplification, Figure 5 showcases selected examples extracted from these datasets, offering a glimpse into the diversity of traffic sign images utilized in this study.

Figure 5. Some pictures in CCTSDB and GTSDB.

4.2. Evaluation Criteria

In this paper, the performance of the algorithm model will be evaluated using three metrics, AP, AP50, and AP75, from the COCO dataset, where AP50 is the average precision obtained when the detector threshold is greater than 50, and AP75 is the average precision obtained when the detector threshold exceeds 75. The calculation of AP in the COCO dataset is based on the precision–recall curve. First, for each category, the predictions are ranked according to their confidence level, and then the true positives (TP) and false positives (FP) are calculated for each prediction. Then, precision and recall are calculated based on TP and FP, and a precision–recall curve is plotted. Finally, the area under curve (AUC) is calculated, which is the AP value of the category. The final AP value of the model is obtained by averaging the AP values of all categories.

5. Experiments

5.1. Experimental Details

The experiments detailed in this paper were conducted using the PyTorch deep learning framework on a 64-bit Linux system, utilizing an NVIDIA GeForce RTX3090 (made by NVIDIA in Santa Clara, CA, USA) graphics card with 24 GB of video memory. During the training phase, the datasets were divided into training, test, and validation sets in an 8:1:1 ratio. The images were resized to 800 × 800 pixels.

The learning rate is a critical parameter that significantly influences the convergence speed of the model. If set too large, it may lead to loss oscillation, while setting it too small may cause the model to converge to a local optimum. After careful consideration, we set the learning rate and weight decay rate to 0.0001 and trained the model for a total of 100 epochs. Comparing the SGD optimizer to the AdamW optimizer, we found that the latter yielded better model convergence. Therefore, we opted to use the AdamW optimizer. The batch size was set to 4, and we utilized 300 query positions.

Data enhancement techniques play a pivotal role in enhancing a model's robustness and preventing overfitting. Hence, we employed various techniques, such as scaling, rotation, and random cropping, during the training process to mitigate overfitting.

5.2. Ablation Study

To assess the efficacy of each component in DSRA-DETR, we conducted ablation experiments on the GTSDB and CCTSDB datasets, evaluating the overall structure of our design as well as the ASPP and FAM modules using AP, AP50, and AP75 as performance metrics. We used Anchor-DETR as a baseline and incrementally improved it with our DSPP and FRAMs, subsequently evaluating its performance on both datasets. The results are shown in Tables 1 and 2.

Table 1. Ablation study for DSRA-DETR.

Settings	GTSDB			CCTSDB		
	AP	**AP50**	**AP75**	**AP**	**AP50**	**AP75**
Baseline	73.61	95.16	90.35	76.92	96.52	94.83
Baseline + (C2.3.4.5)	74.12	96.03	90.87	77.21	97.14	95.24
Baseline + (C2.3.4.5) + DSPP	74.98	96.53	91.12	77.54	97.25	95.95
Baseline + (C2.3.4.5) + DSPP + FRAM	76.13	98.12	92.03	78.24	98.33	97.23

Table 2. Ablation study on DSRA-DETR model regarding multiscale target detection.

Settings	GTSDB			CCTSDB		
	APs	**APm**	**APl**	**APs**	**APm**	**APl**
Baseline	55.82	75.93	85.94	60.23	82.53	93.23
Baseline + (C2.3.4.5)	55.93	76.24	86.01	61.24	82.76	93.55
Baseline + (C2.3.4.5) + DSPP	56.52	76.22	85.93	61.96	82.65	93.41
Baseline + (C2.3.4.5) + DSPP + FRAM	57.12	76.19	86.42	63.04	83.12	94.37

The baseline model achieved an AP score of 73.61% on the GTSDB dataset and 76.92% on the CCTSDB dataset. With the incorporation of the multiscale features, the model's performance was enhanced, resulting in a respective increase of 0.51% and 0.29% in AP scores for the two datasets. Moreover, when we further integrated the DSPP and FRAMs, the model achieved even better results, with improvements of 0.86% and 1.15% in AP scores for the GTSDB dataset, and 0.33% and 0.70% for the CCTSDB dataset, respectively. These results suggest that the proposed DSRA-DETR model can effectively improve the detection performance of traffic signs for both datasets.

Table 2 lists the average accuracy for small targets as APs, medium targets as APm, and large targets as APl. By analyzing the table, we can observe that the model improved the detection performance for all three sizes of targets for both datasets to varying degrees when trained with multilayer features. Specifically, we can see that the detection performance of the model for small targets (AP) was improved to some extent when multilayer features were added. Notably, when DSPP and FRAMs were used for differential aggregation of multiple features, we observed that the AP metrics of the model improved from 55.82% to 57.12% and from 60.23% to 63.04% for the GTSDB and CCTSDB datasets, respectively. This implies that using features extracted from multiple layers and combining the proposed module allow the complex details of the target to be captured and the useless information to be filtered out from the low-level features. The detection performance of the model for small targets was further improved.

In Figure 6, we compare the detection performance of our proposed DSRA-DETR model with the baseline model for the two datasets, CCTSDB and GTSDB. To demonstrate the effectiveness of our model in detecting small targets or multiple small targets, we specifically selected two examples from each dataset for comparative experiments. The results show that our DSRA-DETR model outperforms the baseline model in detecting small targets, which can be attributed to the integration of our proposed FRAM and DSPP modules. Furthermore, to provide further comparative illustration, we present a visual of the feature maps of these examples.

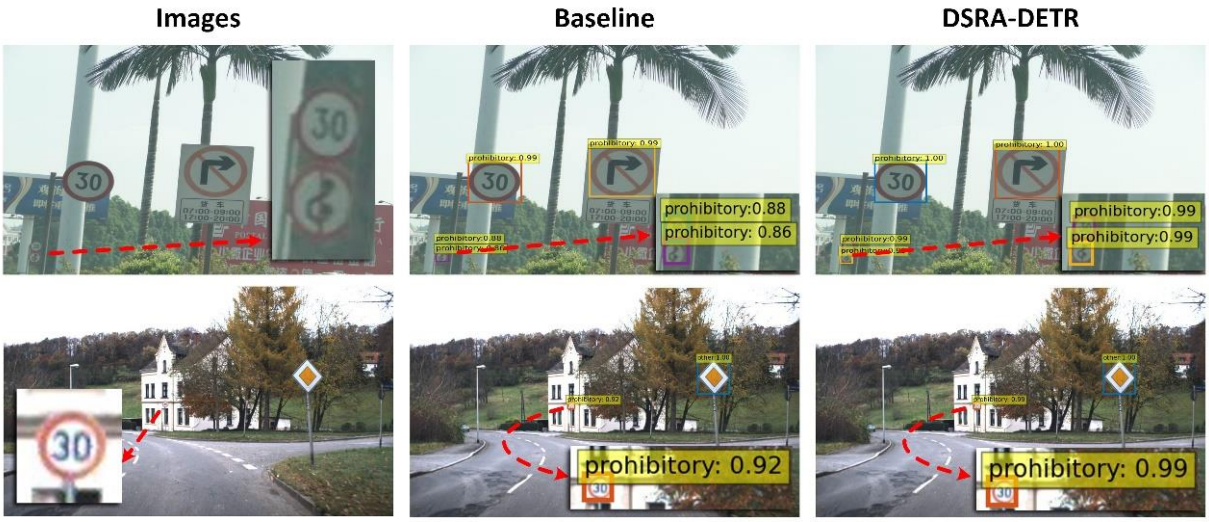

Figure 6. Example of small-target accuracy improvement.

Figure 7 presents visualizations of the feature maps for the two selected exemplary instances, illustrating the effects of the DSPP and FRAMs on enhancing the representation of traffic signs. The feature maps exhibited noticeable improvements with the application of the DSPP module. This module effectively eliminates extraneous information while emphasizing essential aspects such as the spatial location and edge characteristics of the traffic signs. As a result, the feature maps become more focused and discriminative.

Additionally, the FRAM plays a crucial role in augmenting the feature maps' capacity to represent small-scale targets. This enhancement is particularly significant as it enables the model to concentrate more effectively on extracting and leveraging relevant information from small-scale traffic signs during its operational phase. With the incorporation of the FRAM, the model exhibits an improved ability to discern subtle details and capture the distinctive features associated with smaller traffic signs. These visualizations provide compelling evidence of the efficacy of the proposed DSRA-DETR model. The DSPP and FRAMs effectively refine the feature maps, enhancing their representational power and facilitating accurate detection and localization of traffic signs. The combination of these

modules contributes to the overall performance improvements observed in terms of average precision (AP) scores for both the GTSDB and CCTSDB datasets.

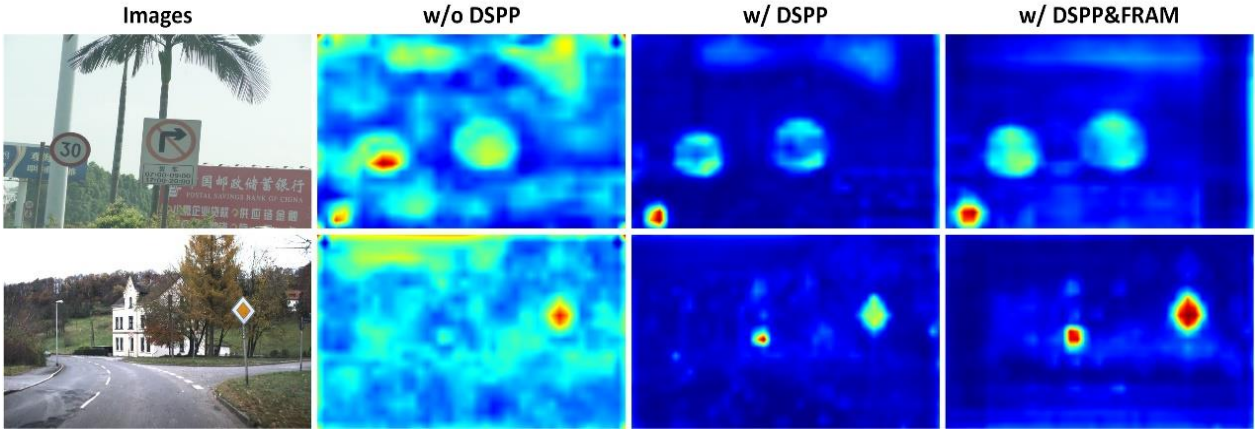

Figure 7. Visualized feature maps.

5.3. Comparison with Previous Methods

In our study, we evaluated the performance of our DSRA-DETR algorithm in comparison to several popular algorithms used for traffic sign detection and general target detection tasks. The algorithms we compared included YOLOv3, Deformable DETR, CornerNet, and Conditional DETR, all of which use the training and evaluation APIs provided by the COCO dataset. We present the results of our experiments in Table 3.

Table 3. Comparison with previous methods.

Method	GTSDB			CCTSDB		
	AP	**AP50**	**AP75**	**AP**	**AP50**	**AP75**
Deformable DETR	73.89	97.15	91.21	75.99	97.18	96.89
CornerNet	56.75	73.84	67.54	57.69	73.52	67.31
YOLOv3	61.28	74.36	76.73	61.92	74.99	73.15
Conditional DETR	73.46	96.85	91.34	77.13	97.21	96.72
DSRA-DETR (Ours)	76.13	98.12	92.03	78.24	98.33	97.23

Compared to Deformable DETR and Conditional DETR, our algorithm achieved a significant improvement in AP for the GTSDB of 2.24% and 2.67%, respectively. Similarly, the APs for the CCTSDB also showed notable enhancements of 2.25% and 1.11%, respectively. These improvements can be attributed to the incorporation of two essential modules, DSPP and FRAM. The DSPP module refines the original feature map by eliminating redundant features and creating a more suitable feature map for the traffic sign detection task. On the other hand, the FRAM aggregates rich location information from lower-level features, leading to better small-target detection performance for higher-level features.

Compared to YOLOv3 and CornerNet, our algorithm demonstrated remarkable improvements in AP for the GTSDB of 14.85% and 19.38%, respectively. For the CCTSDB, the APs were enhanced by 16.32% and 20.55%, respectively. These substantial performance gains can be attributed to the attention mechanism based on the transformer architecture. This novel visual processing method calculates the pixel point's association with other pixel points, offering a different approach from traditional CNN architecture. Moreover, the introduction of the DSPP and FRAMs plays a crucial role in further enhancing the algorithm's overall performance.

The precision–recall curve in Figure 8 clearly shows that our proposed method outperforms all other compared methods, as it has the largest area enclosed by the coordinate

axes. This indicates that our method achieves the best results after training. It is noteworthy that all three methods based on the transformer architecture surpass the performance of the two CNN-based methods, providing further confirmation of the effectiveness of the transformer architecture.

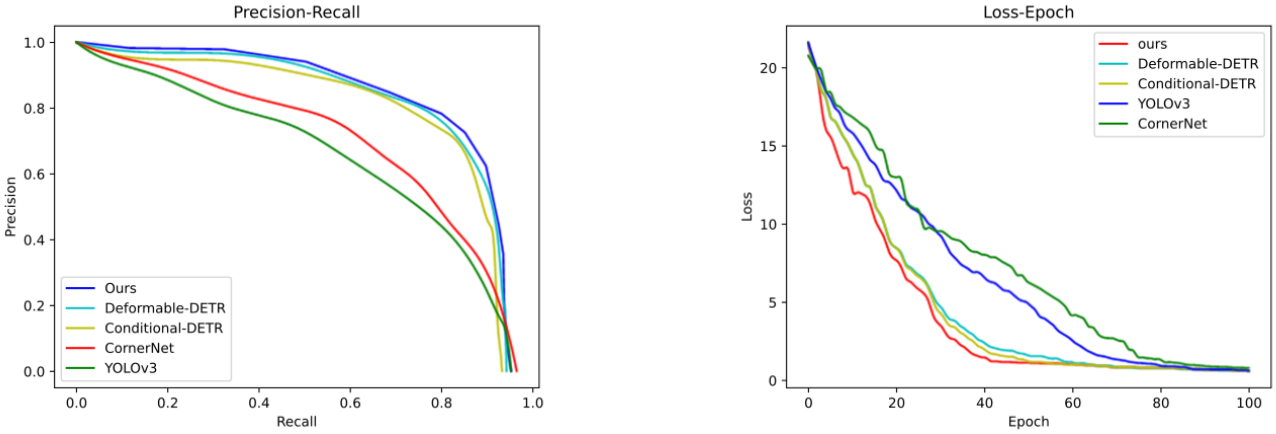

Figure 8. The precision–recall curve and loss–epoch curve.

The loss–epoch curves depicted in Figure 8 demonstrate that our proposed method exhibits superior convergence speed compared to other methods. This can be attributed to Deformable-DETR, Conditional-DETR, and Anchor-DETR, which accelerate convergence. However, it is important to acknowledge that transformer-based algorithms typically require more time to complete an epoch than CNN-based algorithms.

6. Conclusions and Discussion

In this paper, we introduce DSRA-DETR, a novel method for multiscale traffic sign detection. Our method incorporates an efficient feature fusion module to enhance Anchor-DETR. Unlike traditional CNN-based detectors, we leverage the transformer architecture, which has shown great potential in various computer vision tasks. We investigated different feature fusion methods and pyramidal feature map generation and found that integrating multilevel feature maps maximizes their effectiveness in traffic sign detection. Additionally, we integrated the DSPP module to enhance feature information and improved localization capability at each level. Moreover, the FRAM was employed for feature aggregation, enabling our model to capture valuable underlying feature information and further enhance performance.

Extensive experiments on the GTSDB and CCTSDB datasets demonstrate that DSRA-DETR outperforms several advanced target detection methods in terms of accuracy. However, it is important to acknowledge that the transformer-based model requires significant memory and computational power, making it challenging to deploy in real self-driving vehicle systems for sustainable urban life. In future research, we propose focusing on lightweighting and real-time optimization, aiming to reduce model size and computational requirements without compromising accuracy. This would be beneficial for improving the algorithm and its applicability in sustainable urban living.

In conclusion, our proposed DSRA-DETR method offers a promising solution for multiscale traffic sign detection, showcasing its effectiveness and surpassing existing methods in accuracy. In future research, we aim to explore a lightweight and real-time traffic sign detection algorithm suitable for deployment in autonomous vehicle systems to enhance road safety. Furthermore, we aim to promote the application of artificial intelligence in sustainable urban living, contributing to a safer and more efficient traffic management system.

Author Contributions: Conceptualization, J.X.; Funding acquisition, X.C.; Investigation, J.X. and W.L.; Supervision, M.L. and X.C.; Writing—original draft, J.X.; Writing—review and editing, M.L. and X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research reported herein was supported by the NSFC of China under Grant No. 71571091 and 71771112.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All results and data obtained can be found in open access publications.

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

References

1. Anjum, M.; Shahab, S.J.S. Emergency Vehicle Driving Assistance System Using Recurrent Neural Network with Navigational Data Processing Method. *Sustainability* **2023**, *15*, 3069. [CrossRef]
2. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
3. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
4. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37. [CrossRef]
5. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.J.A.P.A. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
6. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional detr for fast training convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3651–3660. [CrossRef]
7. Wang, Y.; Zhang, X.; Yang, T.; Sun, J. Anchor detr: Query design for transformer-based detector. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2567–2575. [CrossRef]
8. Houben, S.; Stallkamp, J.; Salmen, J.; Schlipsing, M.; Igel, C. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 4–9 August 2013; pp. 1–8. [CrossRef]
9. Zhang, J.; Zou, X.; Kuang, L.-D.; Wang, J.; Sherratt, R.S.; Yu, X.J.H.-c.C.; Sciences, I. CCTSDB 2021: A more comprehensive traffic sign detection benchmark. *Hum.-Centric Comput.* **2022**, *12*, 23. [CrossRef]
10. Redmon, J.; Farhadi, A. Yolov3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
12. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
13. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Proc. Syst.* **2015**, *28*, 1137–1149. [CrossRef] [PubMed]
14. 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. [CrossRef]
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969. [CrossRef]
16. Tabernik, D.; Skočaj, D. Deep learning for large-scale traffic-sign detection and recognition. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1427–1440. [CrossRef]
17. Wang, F.; Li, Y.; Wei, Y.; Dong, H. Improved faster rcnn for traffic sign detection. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; pp. 1–6. [CrossRef]
18. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271. [CrossRef]
19. Bochkovskiy, A.; Wang, C.-Y.; Liao, H. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
20. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13039–13048. [CrossRef]

21. Fu, C.-Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659. [CrossRef]
22. Li, Z.; Zhou, F. FSSD: Feature fusion single shot multibox detector. *arXiv* **2017**, arXiv:1712.00960. [CrossRef]
23. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212. [CrossRef]
24. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
25. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750. [CrossRef]
26. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578. [CrossRef]
27. Tian, Y.; Gelernter, J.; Wang, X.; Li, J.; Yu, Y. Traffic sign detection using a multi-scale recurrent attention network. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 4466–4475. [CrossRef]
28. Wang, J.; Chen, Y.; Dong, Z.; Gao, M. Applications. Improved YOLOv5 network for real-time multi-scale traffic sign detection. *Neural Comput. Appl.* **2022**, *35*, 7853–7865. [CrossRef]
29. Zou, H.; Zhan, H.; Zhang, L. Neural Network Based on Multi-Scale Saliency Fusion for Traffic Signs Detection. *Sustainability* **2022**, *14*, 16491. [CrossRef]
30. Xiao, G.; Luo, H.; Zeng, K.; Wei, L.; Ma, J. Robust Feature Matching for Remote Sensing Image Registration via Guided Hyperplane Fitting. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 5600714. [CrossRef]
31. Xiao, G.; Ma, J.; Wang, S.; Chen, C. Deterministic Model Fitting by Local-Neighbor Preservation and Global-Residual Optimization. *IEEE Trans. Image Process.* **2020**, *29*, 8988–9001. [CrossRef]
32. Malik, Z.; Siddiqi, I. Detection and recognition of traffic signs from road scene images. In Proceedings of the 2014 12th International Conference on Frontiers of Information Technology, Islamabad, Pakistan, 17–19 December 2014; pp. 330–335. [CrossRef]
33. Tang, S.; Huang, L.-L. Traffic sign recognition using complementary features. In Proceedings of the 2013 2nd IAPR Asian Conference on Pattern Recognition, Naha, Japan, 5–8 November 2013; pp. 210–214. [CrossRef]
34. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 213–229. [CrossRef]
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Proc. Syst.* **2017**, *30*, 5998–6008. [CrossRef]
36. 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; pp. 740–755. [CrossRef]
37. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159. [CrossRef]
38. Li, F.; Zeng, A.; Liu, S.; Zhang, H.; Li, H.; Zhang, L.; Ni, L.M. Lite DETR: An interleaved multi-scale encoder for efficient detr. *arXiv* **2023**, arXiv:2303.07335. [CrossRef]
39. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13619–13627. [CrossRef]
40. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605. [CrossRef]
41. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]