World Electric Vehicle Journal

MDPI

Article

TF-YOLO: A Transformer–Fusion-Based YOLO Detector for Multimodal Pedestrian Detection in Autonomous Driving Scenes

Yunfan Chen, Jinxing Ye and Xiangkui Wan *

Hubei Key Laboratory for High-Efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan 430068, China; yfchen@hbut.edu.cn (Y.C.); 102210252@hbut.edu.cn (J.Y.)
* Correspondence: xkwan@hbut.edu.cn

Abstract: Recent research demonstrates that the fusion of multimodal images can improve the performance of pedestrian detectors under low-illumination environments. However, existing multimodal pedestrian detectors cannot adapt to the variability of environmental illumination. When the lighting conditions of the application environment do not match the experimental data illumination conditions, the detection performance is likely to be stuck significantly. To resolve this problem, we propose a novel transformer–fusion-based YOLO detector to detect pedestrians under various illumination environments, such as nighttime, smog, and heavy rain. Specifically, we develop a novel transformer–fusion module embedded in a two-stream backbone network to robustly integrate the latent interactions between multimodal images (visible and infrared images). This enables the multimodal pedestrian detector to adapt to changing illumination conditions. Experimental results on two well-known datasets demonstrate that the proposed approach exhibits superior performance. The proposed TF-YOLO drastically improves the average precision of the state-of-the-art approach by 3.3% and reduces the miss rate of the state-of-the-art approach by about 6% on the challenging multi-scenario multi-modality dataset.

Keywords: deep learning; convolutional neural network; multimodal images; pedestrian detection

Citation: Chen, Y.; Ye, J.; Wan, X. TF-YOLO: A Transformer–Fusion-Based YOLO Detector for Multimodal Pedestrian Detection in Autonomous Driving Scenes. *World Electr. Veh. J.* **2023**, *14*, 352. https://doi.org/10.3390/wevj14120352

Academic Editor: Ghanim A. Putrus

Received: 15 November 2023
Revised: 2 December 2023
Accepted: 15 December 2023
Published: 18 December 2023

1. Introduction

Researchers in the field of computer vision have always been interested in pedestrian detection since it is a crucial approach for many applications, including intelligent robotics, naturalistic driving, autonomous driving, and intelligent transportation systems (ITSs) [1–4]. For good illumination conditions, desirable pedestrian detection performance can be achieved using many existing methods [1] which use a visible (VI) image as input. In contrast, it is quite challenging to achieve robust performance for adverse illumination conditions, such as nighttime, low light, total darkness, shadows, and overexposure, as shown in Figure 1. Since the VI cameras rely on good lighting conditions in the environment, they tend to capture low-quality images when lighting is poor. According to the National Highway Traffic Safety Administration (NHTSA) Fatality Analysis Reporting System (FARS), the majority of pedestrian fatalities due to traffic accidents occur in harsh lighting scenarios, such as nighttime and bad weather [5]. Therefore, this work focuses on improving the effectiveness of pedestrian detectors in low-light conditions. Combining multispectral images (visible and infrared images) has proven useful for robust pedestrian detection in ADAS applications [6–21]. The signals of visible and infrared images originate from different modes and can provide scene information from different aspects. Visible images capture reflected light, while infrared images capture thermal radiation. Therefore, this combination is more informative than the combination of single-modal signals. Visible images typically have high spatial resolution, considerable detail, and light–dark contrast. Therefore, they are suitable for human visual perception. However, these images are easily

affected by harsh conditions such as poor lighting, fog, and other adverse weather effects, as shown in Figure 1. Infrared images are resistant to these interferences but typically have low resolution and poor texture. Therefore, the fusion of visible and infrared images can improve the performance of object detection in harsh environmental conditions due to the universality and complementarity of the images utilized. However, it is difficult for multimodal pedestrian detectors to adapt to the changing environment. The detection performance is likely to be stuck when the actual application illumination conditions are different from the illumination conditions of the experiment data.

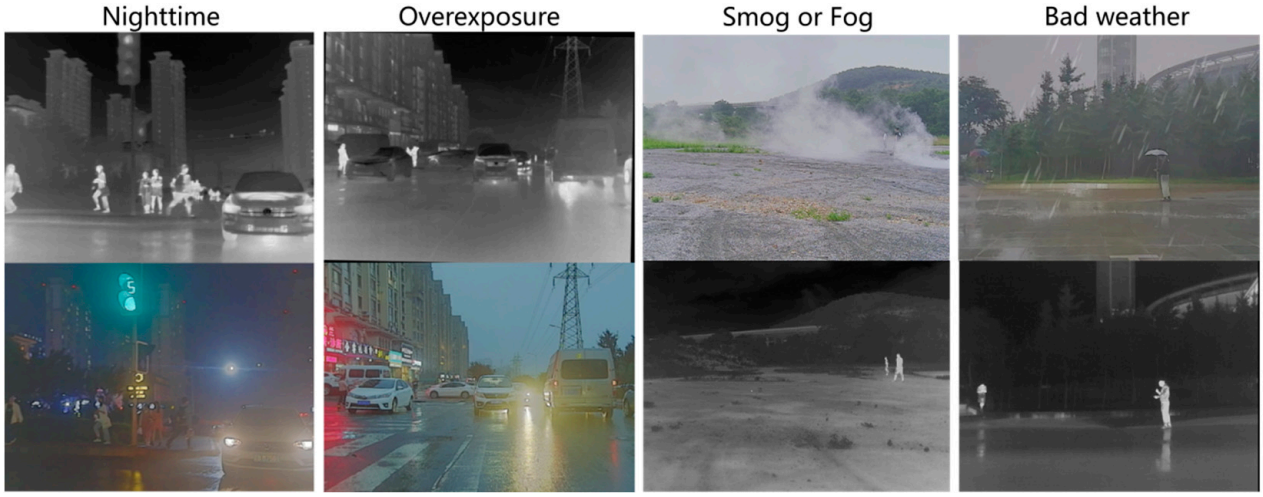


Figure 1. Sample images to show the difficult cases for multimodal pedestrian detection. The top are visible images and the bottom are infrared images.

Therefore, we develop a transformer–fusion-based YOLO detector to effectively combine the multimodal images captured by car-mounted visible and infrared cameras which aim to detect pedestrians regardless of adverse illumination environments. The main contributions of this work are summarized as follows.

- A novel transformer–fusion-based YOLO (TF-YOLO) is introduced to effectively fuse the visible and infrared images for multimodal pedestrian detection, enabling precise pedestrian detection in low-light conditions.
- In our TF-YOLO, we first design a two-stream backbone of YOLOv7 [22] to extract multimodal features. Then, we develop a transformer–fusion module to fuse the input visible and infrared data in the two-stream feature extraction backbone in several positions. It can efficiently combine the rich semantic features in high-level and high-resolution detailed features at a low level, which deeply exploits the long-range multi-modal information.
- Our method achieves the best performance with an average precision of 87.85% and a miss rate of 17.27%, which achieves a 5.1% improvement in average precision and a 6.05% improvement in miss rate, respectively, when compared with the state-of-the-art CFT method [23], on the multi-scenario multi-modality dataset (M^3FD) [24].

The remainder of this article is organized as follows. Section 2 reviews related studies in multimodal pedestrian detection. In Section 3, we present our proposed method in detail. Section 4 is dedicated to the presentation of results and subsequent discussion. To conclude, our work is summarized in Section 5.

2. Related Works

Multispectral imagery-related research has experienced a boom over the past decade, especially for pedestrian detection in Advanced Driving Assistance System applications by combining the different bands (visible and infrared) captured. Multispectral pedestrian detection aims to improve pedestrian detection that is resistant to variations in illumination

and occlusion by utilizing the complementary information of multi-modal data (visible and infrared). Hwang et al. [6] developed a benchmark and enhanced aggregated channel features (ACF) to facilitate multispectral pedestrian detection. In recent years, several multispectral pedestrian detection techniques based on deep convolutional neural networks (CNNs) have been presented, owing to the rapid growth of CNNs [7–21]. In [7], a CNN is first adopted for multispectral pedestrian detection, and a two-stream CNN with the late fusion strategy is proposed. Liu et al. [8] explored several fusion strategies, early fusion, halfway fusion, and late fusion, on Faster R-CNN [9]. They found that halfway fusion performs best. Chen et al. [10] found that multi-layer fusion performs better than half-way fusion and proposed a multilayer fusion technique. Li et al. [11] discovered that illumination variance has effects on detection confidence. Motivated by this, they carried out a weighted score fusion by illumination score and suggested an illumination-aware subnet to forecast the illumination conditions of the input images. A differential modality-aware fusion module for complementary fusion was introduced in [12] and conducts channel-wise differential weighting. To balance the detection accuracy and fusion performance, a deconvolutional single-shot detector with a multilayer fusion method is given in [13]. An adaptive method for fusing multi-modal data was created by Zhang et al. [14] using guided attentive feature fusion.

To achieve better multispectral pedestrian identification, various systems investigate the addition of an auxiliary job to multi-modal feature fusion. To concurrently detect and segment pedestrians, a CNN network is proposed in [15]. Real-time multispectral pedestrian identification could benefit from the addition of box-level segmentation supervision, as suggested by Cao et al. [16]. Zhang et al. [17] created a knowledge distillation network that uses a teacher network with high-resolution feature fusion to instruct a student network with low-resolution image-level fusion to overcome the hardware and software constraints of multispectral pedestrian identification. In [18], a lightweight anchor-free method based on local and global hybrid attention mechanisms is developed for multispectral pedestrian detection. Some other methods [19–21] were explored to solve the alignment problem.

Although the above methods have made great contributions to the progress of multispectral object detection, multispectral pedestrian detection under environments with changing illuminations still faces unresolved limitations. Therefore, the objective of this work is to explore generalizable multispectral pedestrian detection to improve the accuracy and effectiveness of pedestrian detection.

3. Proposed Methods

3.1. Overview

The framework of TF-YOLO is shown in Figure 2. To demonstrate the effectiveness of our developed transformer–fusion module, we redesign the framework of YOLOv7 [22] to enable multimodal pedestrian detection. The backbone of TF-YOLO consists of a two-stream feature extraction network and three transformer–fusion modules. The head of TF-YOLO has three outputs: Y1, Y2, and Y3. To extract multimodal information, the two-stream backbone processes the incoming visible and infrared pictures first. Then, the extracted multimodal features are integrated by the proposed transformer–fusion modules at three positions of the backbone. Finally, the head network outputs the detection results of pedestrians.

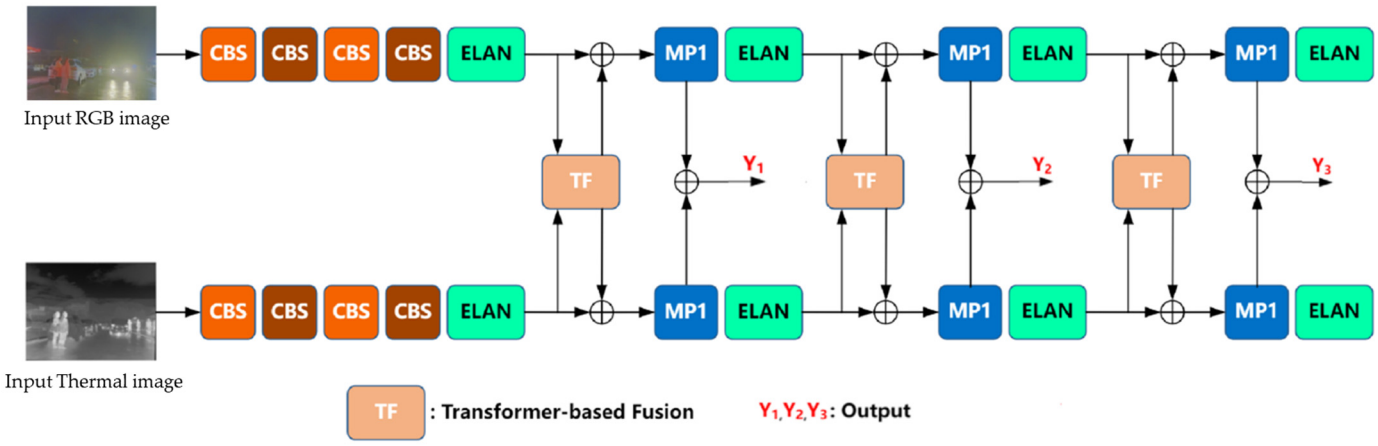


Figure 2. Framework of the proposed TF-YOLO. The backbone of TF-YOLO includes a two-stream feature extraction network and three transformer–fusion modules. The head of TF-YOLO has three outputs: Y1, Y2, and Y3.

3.2. Transformer–Fusion-Based Two-Stream Backbone

The adverse illumination environment usually has nighttime, dense smoke, dust, bad weather, etc., which are likely to make the pedestrian poorly visible. The effectiveness of pedestrian detection can be enhanced by using multispectral data, which consists of both visible and infrared images. Existing multispectral feature fusion strategies mainly include summation fusion, concatenation fusion, and illumination-aware fusion. However, These fusion algorithms are primarily centered on local characteristics, which are neither resistant to the complexity and variety of the environment in autonomous driving situations, nor do they adequately utilize long-range contextual information in both intra-modality and cross-modality. Recent research [23,25] has demonstrated the usefulness of transformers in representing long-range relationships when compared to convolutional neural networks (CNNs). This encourages us to use a transformer to take advantage of the distant contextual relationships between various input items. The self-attention mechanism in the transformer uses axial attention to model long-range dependencies, which is beneficial to the network learning global contextual features. When performing two-stream network feature fusion, local and global context information can be enhanced at the same time, maximizing the fusion of infrared and visible light information while avoiding information loss. Thus, we develop a transformer–fusion module to deeply integrate the visible and infrared information. The reason why the proposed fusion module can adapt to changes in environmental illumination is that infrared and visible light information are adaptively fused through learning. The module can dynamically adapt to the fusion according to changes in input information. Unlike fixed fusion rules, some important information will be lost due to changes in input information.

Figure 3 displays the architecture of our developed transformer–fusion module. By using the RGB feature map F_{VI} and the thermal feature map F_{IR} as inputs, the proposed transformer–fusion module creates the multi-modal fusion feature map F_M. We first perform 3-D attention weights for the input feature maps to enhance the feature representation capabilities. In particular, the inputs F_{VI} and F_{IR} are fused with themselves via a dot product after passing through 3-D weights, respectively. Subsequently, the flattened features F'_{VI} and F'_{IR} are concatenated along the channel dimension to produce the feature vector F'_C. After that, a learnable positional embedding is added with F'_C to generate the input which will then be sent to the transformer blocks. The positional embedding can encode the position information into F'_C which helps differentiate spatial information between different tokens at training time. The transformer module consists of 8 transformer blocks,

each with two normalization layers, a multi-head attention mechanism, and an MLP which consists of a two-layer fully connected feed-forward network with a GELU activation [26].

$$T' = MultiHead(Q, K, V) = Concat(T_1, \ldots, T_h)\omega^O \quad (1)$$

$$T_i = Attention\left(Q\omega_i^Q, K\omega_i^K, V\omega_i^V\right) \quad (2)$$

where the subscript h means the number of heads, and ω^O indicates the projected matrix of $Concat\ (T_1, \ldots, T_h)$. Q, K, V are three-input weight matrices of multi-head attention which are projected by I. ω^O, ω^O, and ω^O are weight matrices. The self-attention layer computes the weights of attention using scaled dot products between $Q\omega_i^Q, K\omega_i^K, V\omega_i^V$ and it is then multiplied by the values to infer the refined output T_i.

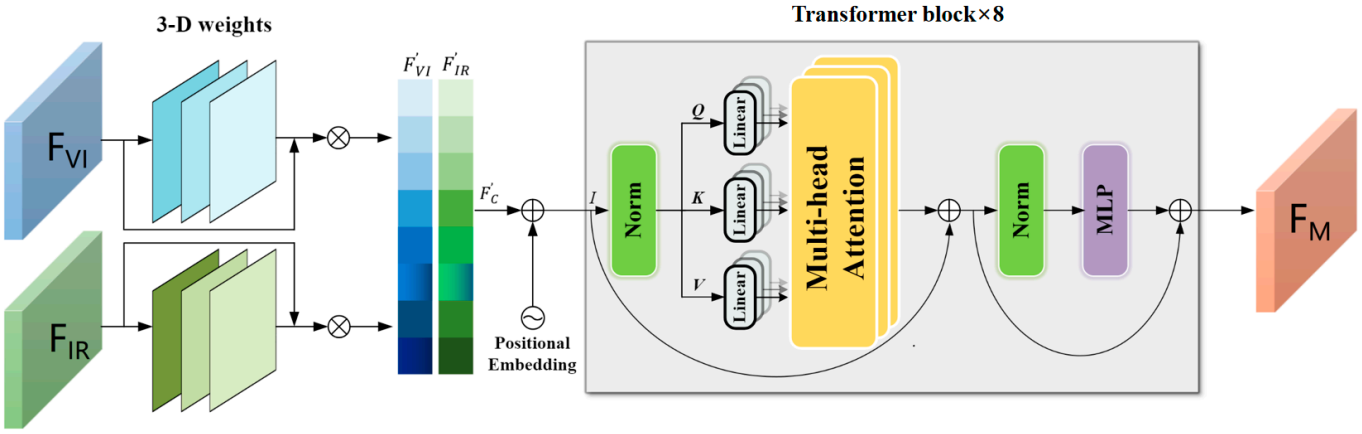


Figure 3. Transformer–fusion module.

3.3. Training

3.3.1. Loss Function

A summation of the regression loss (L_{bbox}) of the bounding box, the classification loss (L_{cls}), and the confidence loss (L_{conf}) forms the overall loss function,

$$L = L_{bbox} + L_{cls} + L_{conf} \quad (3)$$

The L_{bbox}, L_{cls}, and L_{conf} are calculated by the following functions, where a Generalized Intersection over Union ($GIoU$) loss [27] is adopted to calculate the L_{bbox}.

$$L_{bbox} = \sum_{i=0}^{S^2}\sum_{j=0}^{N} C_{i,j}^{obj} \cdot L_{GIoU_i} = \sum_{i=0}^{S^2}\sum_{j=0}^{N} C_{i,j}^{obj} \cdot [1 - GIoU_i] = \sum_{i=0}^{S^2}\sum_{j=0}^{N} C_{i,j}^{obj} \cdot \left[1 - \frac{B_i^g \cap B_i^p}{B_i^g \cup B_i^p} + \frac{B_i^c \backslash (B_i^g \cup B_i^p)}{B_i^c}\right] \quad (4)$$

$$L_{cls} = \sum_{i=0}^{S^2}\sum_{j=0}^{N} C_{i,j}^{obj} \cdot \sum_{ceclasses} p_i(c) log\left(\hat{p_i}(c)\right) \quad (5)$$

$$L_{conf} = \sum_{i=0}^{S^2}\sum_{j=0}^{N} C_{i,j}^{obj} \cdot \left(c_i - \hat{c_i}\right)^2 + \sum_{i=0}^{S^2}\sum_{j=0}^{N} C_{i,j}^{noobj} \cdot \left(c_i - \hat{c_i}\right)^2 \quad (6)$$

where S^2 and N represent the quantity of image grids and prediction boxes, respectively, in each grid throughout the prediction process. B^g, B^p, and B^c are the ground truth, the prediction box, and the smallest closed box surrounding B^g and B^p, respectively. The coefficient $C_{i,j}^{obj}$ represents whether the jth prediction box of the ith grid is a positive sample. The classification loss L_{cls} takes the form of cross-entropy, $p_i(c)$ denotes the probability of the real sample belonging to class c, and $\hat{p_i}(c)$ denotes the probability of the network predicted sample belonging to class c. The final confidence loss contains two units, both squared error losses. The definition of the parameter $C_{i,j}^{noobj}$ is opposite to the coefficient

$C_{i,j}^{obj}$. Finally, c and $\hat{c}$ represent the true confidence value and the network-predicted confidence value.

3.3.2. Training Details

The network is optimized during training using a stochastic gradient descent (SGD) algorithm, with the initial learning rate set to 10^{-2}, the momentum set to 0.937, and the weight decay set to 0.0005. All experiments are conducted on a machine under Ubuntu 18.04 with an Intel Xeon (R) Gold 5218R 2.1 GHz central processing unit (CPU), 64 GB random-access memory, and one NVIDIA GeForce RTX 3090 GPU. Inspired by the Mosaic method [28], our model is trained for 100 epochs with a batch size of 4, and we use a pre-trained YOLOv7 model on the COCO dataset [29] as the initial weight. Our programs are based on Pytorch, which necessitates the use of the CUDA deep neural network library and compute unified device architecture (cuDNN).

4. Results and Discussions

4.1. Datasets

4.1.1. Multi-Scenario Multi-Modality Dataset (M^3FD)

The multi-scenario multi-modality dataset (M^3FD) [24] includes 4200 pairs of aligned visible and infrared images for fusion-based detection tasks. The synchronized system for capturing visible and infrared images contains one binocular optical camera (1024 × 768) and one binocular infrared sensor (640 × 512 before alignment). The synchronized system calibrates the visible images while the homography matrix is added to distort the infrared images artificially. The M^3FD was captured under various lighting scenarios, i.e., Daytime, Overcast, Night, and Challenge, with ten sub-scenarios. It should be emphasized that the shooting scenes of this dataset contain many overexposed scenes caused by car lights or traffic and scenes containing a large amount of smoke, making object detection more challenging. In addition, the dataset contains pedestrians of various scales and a variety of postures, which are suitable for surveillance and autonomous driving. A large amount of data and the diversity of the M^3FD dataset provide convenience for the training and verification of target detection tasks based on multi-sensor fusion. We separated the entire dataset into sections for testing and training. There are 3360 pairs of photos in the training section, 420 pairs in the validation phase, and 420 pairs in the testing part. Labels and bounding box coordinates are present in the ground truth.

4.1.2. UTokyo Multispectral Object Detection Dataset

The UTokyo dataset [30] contains a total of 3740 sets of images taken during the day and 3772 sets of images collected at night, including 1446 sets of pixel-level aligned infrared and visible images, with a resolution of 320 × 256. The dataset was collected at a rate of one frame per second using visible, far-infrared, mid-infrared, and near-infrared sensors and contains five labeled categories, namely bike, car, car_stop, color_cone, and person. This dataset contains a large number of nighttime scenes. In addition to nighttime scenes with street lights or car lights, it also includes a large number of almost completely dark scenes, which are very challenging. Therefore, this dataset is suitable for evaluating object detection methods based on multispectral sensor fusion. In this paper, we use 1446 sets of aligned infrared and visible images for testing to evaluate the performance of pedestrian detection.

4.2. Evaluation Metrics

4.2.1. Precision and Recall

The precision–recall curve is a frequently employed statistic for assessing object detection techniques. True positive (TP), false positive (FP), and false negative (FN) are the three groups into which detection results are classified based on the overlap between anticipated bounding boxes and ground truth boxes. TP stands for pedestrians that are accurately anticipated. A projected bounding box is usually regarded as a TP if the overlap ratio between it and the ground truth is more than 0.5. Whereas FP accounts for non-pedestrian

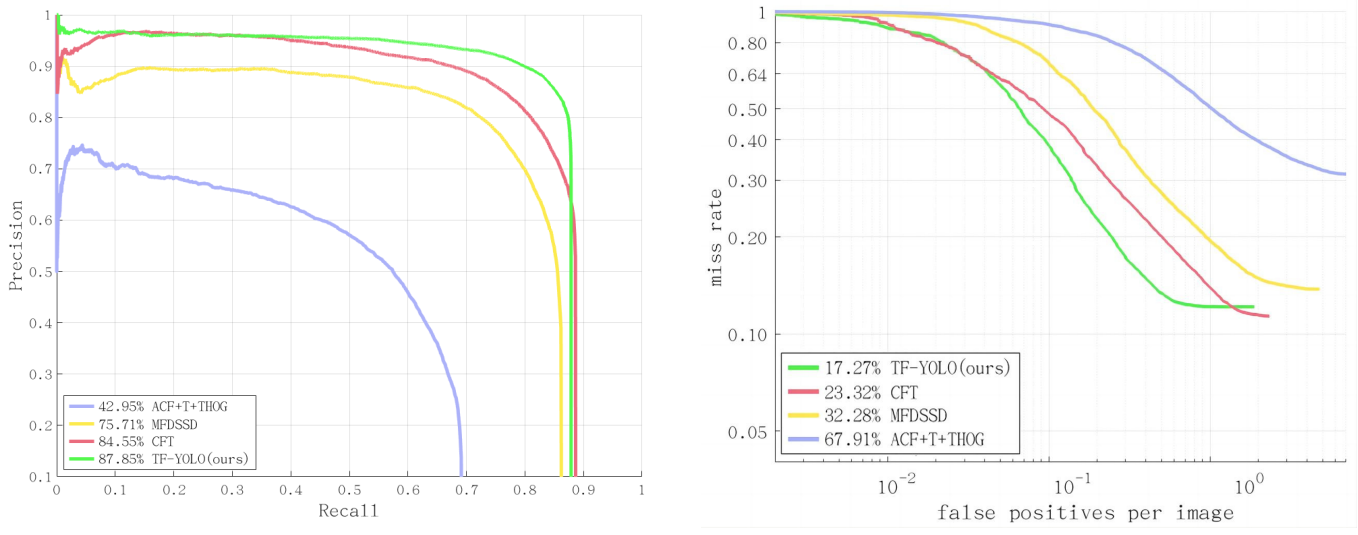
regions that are mistakenly designated as pedestrian zones, FN measures the number of missing pedestrians. The definitions of recall and precision are, respectively, TP/(TP + FP) and TP/(TP + FN). By varying the confidence score threshold, accuracy scores at uniformly spaced recall levels are averaged to determine the average precision (AP). To compute AP, we split the recall levels evenly between 0 and 1 and set the AP value to 100.

4.2.2. Log-Average Miss Rate (MR)

The log-average miss rate (MR) versus a false positive per image (FPPI) range of (10^{-2}, 100) is also used for assessing the performance of the presented method. MR describes the index of detecting the missed detection rate in the detection results, which is calculated by 1-recall; the lower the value of the miss rate, the better. FPPI describes the average false detection rate per image. Suppose there are N images and the number of false checks in the result is FP, then FPPI is calculated by FP/N. MR-FPPI is similar to precision–recall. They are two mutually exclusive indicators. The improvement of one performance will inevitably decrease the other performance, which can reflect the overall performance of the detector. As commonly used settings, the detected bounding box and the ground truth bounding box are matched by selecting a 0.5 minimum overlap ratio.

4.3. Detection Evaluation on the M^{3}FD

The performance of the proposed TF-YOLO is evaluated by comparing it with three other approaches, including ACF + T + THOG [6], MFDSSD [13], and CFT [24]. Figure 4 presents the precision–recall curves and miss rate versus FPPI curves, respectively. It can be observed that our approach obviously outperforms all other methods and achieves the highest AP of 86.56% and the lowest MR of 17.27%, which evidently outperforms the state-of-the-art results of CFT. Furthermore, the performance gap is quite large when compared with MFDSSD [13] and ACF + T + THOG, with 87.85% AP of ours versus 75.71% AP of the MFDSSD and 42.95% AP of the ACF + T + THOG, respectively. The proposed method is dramatically superior to all other methods, which demonstrates that the proposed transformer–fusion module can significantly improve the accuracy of pedestrian detection with bad illuminations. The results of our method outperform all other techniques, which makes sense because the M^{3}FD is taken under environments with low visibility and bad weather, and the proposed transformer–fusion mechanism can dramatically increase the detection accuracy of different illumination and weather conditions.



(**a**) (**b**)

Figure 4. Comparison of precision–recall curves and MR-FPPI curves on the M^{3}FD dataset. (**a**) Precision–recall curves; (**b**) MR−FPPI curves.

Figure 5 displays a comparison of visual detection results for three example images covering challenging scenes including people of varying proportions, low illumination, and smog. Visual comparison clearly shows that our method outperforms all other methods. The state-of-the-art methods produce many false negatives and false alarms. However, our method successfully detects pedestrians at different scenes. One can see that the CFT and MFDSSD methods falsely detected pedestrians and missed pedestrians, while the ACF + T + THOG method produces many false alarms. We can conclude the proposed method is effective in pedestrian detection from different illumination and weather conditions. This result demonstrates that our transformer–fusion module can robustly fuse multispectral data.

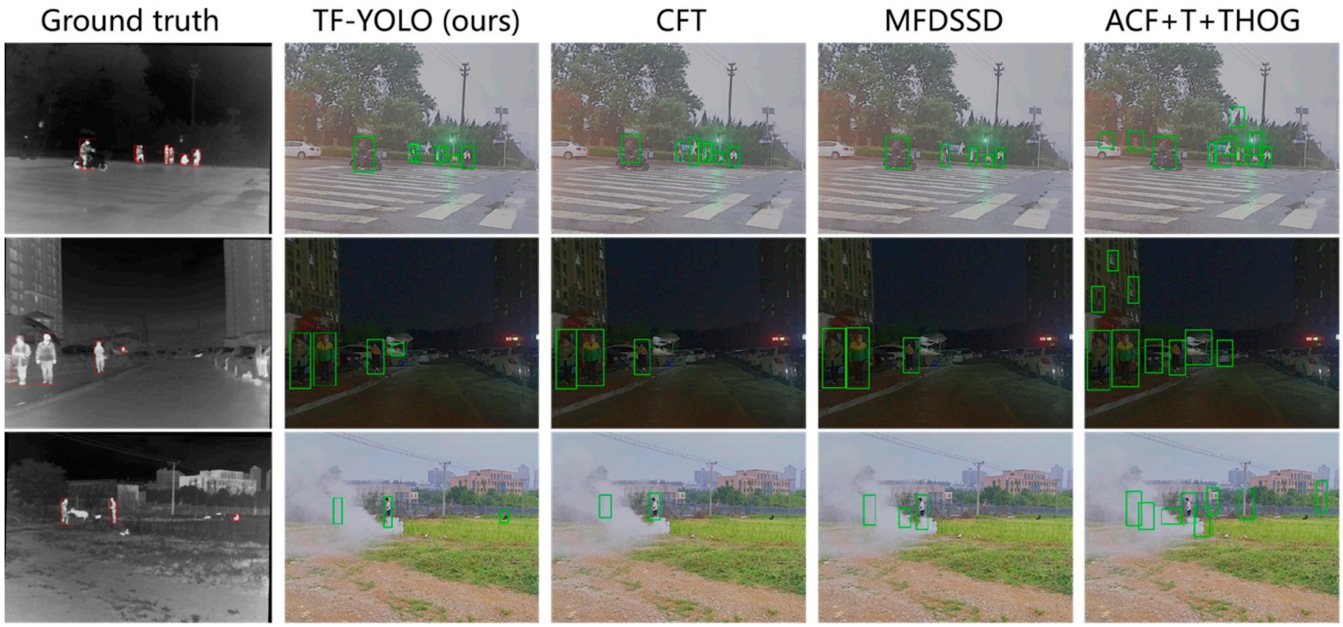
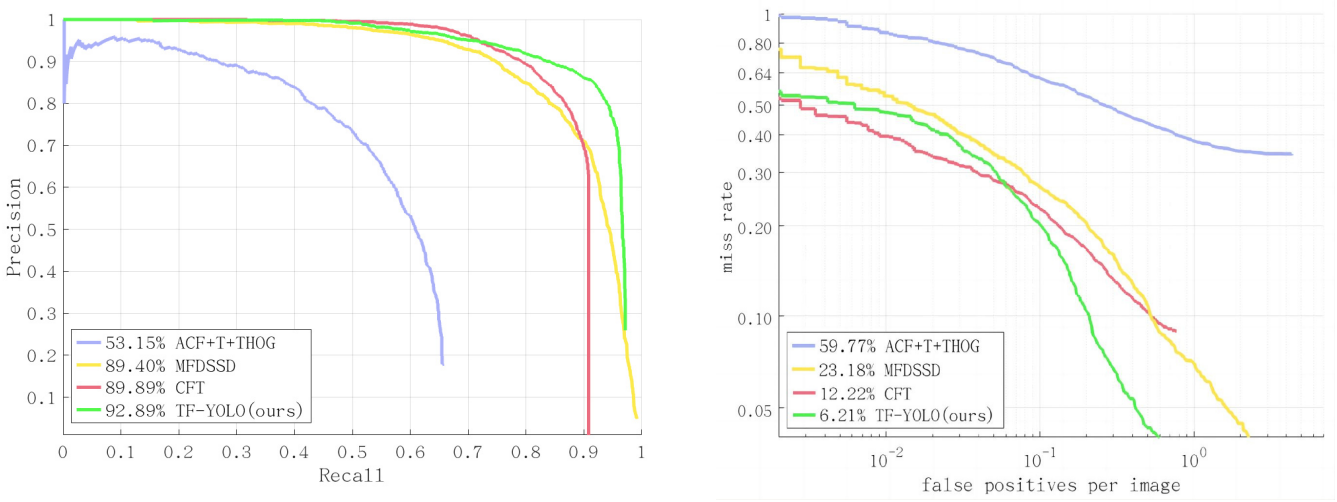


Figure 5. Visual comparison between different approaches and our detection results in the M^3FD dataset. The top to bottom shows the results on three different scenes, daytime, nighttime, and smog, respectively.

4.4. Detection Evaluation on the UTokyo Dataset

We additionally assess the developed approach's effectiveness using the Utokyo multispectral dataset. The MR-FPPI and precision–recall curves are displayed in Figure 6, respectively. Our approach performs 3% better in AP and 6.01% better in MR than the CFT method.

Figure 7 shows the visual detection results of all methods. We can see that our method detected all pedestrians correctly while other methods generated many false alarms and missing instances. This detection result is reasonable because the UTokyo multispectral dataset involves the presence of many pedestrians in a dark environment. Our method can effectively fuse visible and thermal images through the proposed fusion module and correctly detect pedestrians under adverse lighting conditions.



(**a**) (**b**)

Figure 6. Comparison of precision–recall curves and MR-FPPI curves on the UTokyo dataset. (**a**) Precision–recall curves, (**b**) MR-FPPI curves.

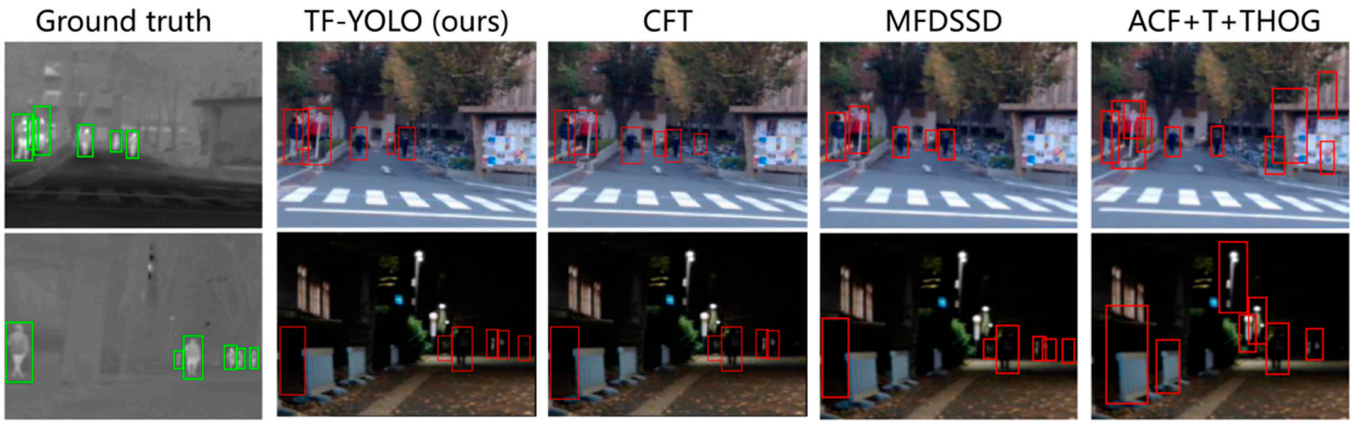


Figure 7. Visual comparison between different approaches and our detection results in the UTokyo dataset. The top to bottom shows the results on two different scenes, daytime and nighttime, respectively.

4.5. Comprehensive Comparison

We make a comprehensive comparison to compare both the detection performance and the average computation time on the M^{3}FD dataset. As shown in Table 1, it is obvious that our TF-YOLO outperforms all other methods in terms of detection performance under all evaluation metrics. Although CFT has the same detection speed (0.05 s/f) as our method, the average precision and miss rate of CFT is quite worse than our method. Therefore, the proposed TF-YOLO has a better trade-off between detection speed and detection performance.

Table 1. Comprehensive comparison on M^{3}FD dataset.

Method	AP (%)	MR (%)	Average Computation Time (s/f)
ACF + T + THOG	42.95	67.91	0.13
MFDSSD	75.71	32.28	0.06
CFT	84.55	23.32	0.05
TF-YOLO (ours)	87.85	17.27	0.05

4.6. Ablation Experiments

In this section, we conduct an ablation study using the M^{3}FD dataset to verify the effectiveness of the proposed TF-YOLO. In the first simulation, we removed the transformer–

fusion module and only retained the two-stream YOLOv7 (fusion by summation) for experiments to check performance. Then, we use visible images and infrared images to train YOLOv7, named YOLOv7-VI and YOLOv7-IR, respectively. This simulation is to evaluate how the proposed two-stream backbone of YOLOv7 contributes to detection performance. Table 2 shows the AP for each simulation.

Table 2. Ablation experimental results on M^3FD dataset.

Method	AP (%)
YOLOv7-VI	75.10
YOLOv7-IR	79.21
Two-stream YOLOv7	82.46
TF-YOLO (Transformer–fusion + Two-stream YOLOv7)	87.85

As shown in Table 2, adding the transformer–fusion module to two-stream YOLOv7 significantly improves the detection performance of AP by 2.77%. It can be seen from the comparison that the combination of infrared and visible branches by the proposed transformer–fusion module is effective in boosting detection performance. In addition, the two-stream YOLOv7 is evidently better than both the YOLOv7-VI and YOLOv7-IR. This result demonstrates that the proposed two-stream backbone of YOLOv7 that fuse infrared and visible images significantly improves detection performance.

4.7. Discussion

4.7.1. Explanation of Results

The effectiveness of the presented TF-YOLO is evaluated by empirical studies, which are covered in Section 4.5. The detection accuracy of TF-YOLO can be significantly increased using the developed transformer–fusion module. We compare our approach with several published state-of-the-art approaches on two well-known datasets in Sections 4.3 and 4.4. We discovered from the experimental findings that our proposed TF-YOLO performs noticeably better than alternative approaches.

In addition to being applied to pedestrian detection in autonomous driving scenarios, the proposed method can also be migrated to object detection in other scenarios, such as surveillance systems, military reconnaissance, drone search, etc. Since applications involving outdoor scenes are often easily affected by the natural environment, such as lighting, weather, and other factors, common object detectors based only on visible cameras have limitations, as mentioned in Section 1. However, the proposed pedestrian detector based on multi-modal sensor fusion overcomes this limitation and has a wider range of application scenarios. The proposed method is scalable and universal. For example, our method can be extended to pedestrian re-identification, animal detection, vehicle detection, saliency detection, etc., by transfer learning.

4.7.2. Limitations of the Proposed Method

As Figure 8 illustrates, our technique demonstrates state-of-the-art detection accuracy and real-time detection speed but fails to perform effectively in some scenarios. Occlusions and extremely rare occurrences are the causes of the failure cases. One person in the crowd is not seen in Figure 8a. In Figure 8b, one person with a small scale is not detected. In Figure 8c, two small-scaled pedestrians occluded by cars are not detected. Thus, pedestrians with occlusions and small scales may reduce detection performance, which is a common issue with most methods. In future works, we will consider exploring depth estimation or a new CNN architecture to resolve these challenges.

(**a**) (**b**) (**c**)

Figure 8. Failure examples of our developed pedestrian detection method. The ground truth is indicated by the red bounding boxes. The detection results are displayed in the green bounding boxes. (**a**–**c**) Failure cases on the selected images from the M^3FD dataset.

5. Conclusions

This paper introduces a new transformer–fusion-based YOLO (TF-YOLO) for accurate multimodal pedestrian detection. Our TF-YOLO contains a transformer–fusion-based two-stream backbone network in which a novel transformer–fusion module is proposed to maximize the complementary characteristics of visible and infrared information to ensure that the detector is not susceptible to interference from the adverse environment, such as adverse illumination and weather conditions. Empirical experiments are conducted on the M^3FD dataset and UTokyo datasets to validate the effectiveness of the proposed method. Our TF-YOLO shows state-of-the-art performance with an 87.85% average precision and a 17.27% miss rate on the M^3FD dataset and a 92.89% average precision and a 6.21% miss rate on the UTokoyo dataset. We anticipate that our findings will help future multispectral pedestrian detection studies.

Author Contributions: Conceptualization, Y.C.; methodology, Y.C., J.Y. and X.W.; software, Y.C.; validation, J.Y.; formal analysis, Y.C. and X.W.; investigation, Y.C. and J.Y.; resources, Y.C.; writing—original draft preparation, Y.C.; writing—review and editing, X.W.; visualization, J.Y.; supervision, X.W.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Hubei Province, China (No. 2023AFB424) and the Open Foundation of Hubei Key Laboratory for High-efficiency Utilization of Solar Energy and Operation Control of Energy Storage System (Project No. HBSEES202314).

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors would like to thank the anonymous reviewers and academic editors for their effort and time invested in the manuscript.

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

References

1. Balsa-Barreiro, J.; Valero-Mora, P.M.; Berné-Valero, J.L.; Varela-García, F.-A. GIS mapping of driving behavior based on naturalistic driving data. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 226. [CrossRef]
2. Balsa-Barreiro, J.; Valero-Mora, P.M.; Menéndez, M.; Mehmood, R. Extraction of naturalistic driving patterns with geographic information systems. *Mob. Netw. Appl.* **2020**, *28*, 619–635. [CrossRef]
3. Chen, L.; Lin, S.; Lu, X.; Cao, D. Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 3234–3246. [CrossRef]
4. Zhang, C.; Berger, C. Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 10279–10301. [CrossRef]
5. Pedestrian Safety: Prevent Pedestrian Crashes. Available online: https://www.nhtsa.gov/road-safety/pedestrian-safety (accessed on 4 October 2021).

6. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; Kweon, I.S. Multispectral Pedestrian Detection: Benchmark Dataset and Baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1037–1045.
7. Wagner, J.; Fischer, V.; Herman, M.; Behnke, S. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. In Proceedings of the 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belgium, 27–29 April 2016; Volume 587, pp. 509–514.
8. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. *arXiv* **2016**, arXiv:1611.02644.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015.
10. Chen, Y.; Xie, H.; Shin, H. Multi-layer fusion techniques using a CNN for multispectral pedestrian detection. *IET Comput. Vis.* **2018**, *12*, 1179–1187. [CrossRef]
11. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171. [CrossRef]
12. Zhou, K.; Chen, L.; Cao, X. Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems. In *Computer Vision–ECCV 2020, Proceedings of the16th European Conference, Glasgow, UK, 23–28 August 2020*; Part XVIII 16; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 787–803.
13. Chen, Y.; Shin, H. Multispectral image fusion based pedestrian detection using a multilayer fused deconvolutional single-shot detector. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **2020**, *37*, 768–779. [CrossRef] [PubMed]
14. Zhang, H.; Fromont, E.; Lefèvre, S.; Avignon, B. Guided Attentive Feature Fusion for Multispectral Pedestrian Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 72–80.
15. Li, C.; Song, D.; Tong, R.; Tang, M. Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv* **2018**, arXiv:1808.04818.
16. Cao, Y.; Guan, D.; Wu, Y.; Yang, J.; Cao, Y.; Yang, M.Y. Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 70–79. [CrossRef]
17. Zhang, H.; Fromont, E.; Lefèvre, S.; Avignon, B. Low-Cost Multispectral Scene Analysis with Modality Distillation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 803–812.
18. Zuo, X.; Wang, Z.; Liu, Y.; Shen, J.; Wang, H. LGADet: Light-weight anchor-free multispectral pedestrian detection with mixed local and global attention. *Neural Process. Lett.* **2023**, *55*, 2935–2952. [CrossRef]
19. Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; Liu, Z. Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5127–5137.
20. Wanchaitanawong, N.; Tanaka, M.; Shibata, T.; Okutomi, M. Multi-Modal Pedestrian Detection with Large Misalignment Based on Modal-Wise Regression and Multi-Modal IoU. In Proceedings of the 2021 17th International Conference on Machine Vision and Applications (MVA), Aichi, Japan, 25–27 July 2021; pp. 1–6.
21. Hu, W.; Fu, C.; Cao, R.; Zang, Y.; Wu, X.-J.; Shen, S.; Gao, X.-Z. Joint dual-stream interaction and multi-scale feature extraction network for multi-spectral pedestrian detection. *Appl. Soft Comput.* **2023**, *147*, 110768. [CrossRef]
22. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
23. Qingyun, F.; Dapeng, H.; Zhaokui, W. Cross-modality fusion transformer for multispectral object detection. *arXiv* **2021**, arXiv:2111.00273.
24. Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-Aware Dual Adversarial Learning and a Multi-Scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5811.
25. Zhang, Y.; Chen, J.; Huang, D. Cat-det: Contrastively Augmented Transformer for Multi-Modal 3d Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 908–917.
26. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
27. Rezatofighi, H.; Tsoi, N.; Gwak, J.Y.; Sadeghian, A. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
28. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.

29. 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 *Computer Vision–ECCV 2014, Proceedings of the13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Part V 13; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
30. Takumi, K.; Watanabe, K.; Ha, Q.; Tejero-De-Pablos, A.; Ushiku, Y.; Harada, T. Multispectral Object Detection for Autonomous Vehicles. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017, New York, NY, USA, 23 October 2017; pp. 35–43.