Scientific Reports

https://doi.org/10.1038/s41598-026-45310-w

Article in Press

Pseudo-depth-based deep neural network model for object detection

Received: 21 January 2026

Accepted: 18 March 2026

Published online: 26 March 2026

Cite this article as: Li S., Feng W., Liu B. *et al.* Pseudo-depth-based deep neural network model for object detection. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-45310-w

Si-Qi Li, Wei Feng, Bin Liu, Xin Tong & Qiang Li

Pseudo-depth-based deep neural network model for object detection

Si-Qi Li [1], Wei Feng [2,3,4], Bin Liu [5], Xin Tong [1], Qiang Li [1†]

[1] *State Key Laboratory of Porous Metal Materials, School of Physical Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China*

[2] *School of Information Mechanics and Sensing Engineering, Xidian University, Xi'an 710071, China*

[3] *Xi'an Key Laboratory of Advanced Remote Sensing, Xi'an 710071, China*

[4] *Shaanxi Innovation Center for Multi-source Fusion Detection and Recognition, China*

[5] *Shanghai Aerospace Control Technology Institute, Shanghai 201109, China*

E-mail: lsq2@mail.nwpu.edu.cn, wfeng@xidian.edu.cn, gray1988@126.com, tongxin@mail.nwpu.edu.cn, liruo@nwpu.edu.cn

Abstract: Current machine learning methods only utilize the three-channel color features of optical images for computer visual tasks. However, the optical images only explicitly present information of RGB color and two-dimensional planar shape, where the third-dimensional spatial features are not fully exploited. This limitation restricts the potential improvement in recognition performance. To address this issue, we propose a detection scheme to enhance model's detection capabilities based on four independent features by combining the pseudo-depth and the RGB features without adding any additional hardware sensors. The monocular depth estimation model is first used as a virtual depth sensor to extract the pseudo-depth features from input optical images. Then the fused Depth-RGB features are fed into the neural network model for object detection training and inference to enhance capability for extracting spatial features. Experiments show that the proposed method has improved the detection metric mAP_{50} by 3.8 and 8.0 percentage points on the public M^3FD and COCO datasets, respectively. Notably, the scheme can be easily embedded into any machine learning models to definitely improve the detection performance.

Keywords: Feature enhancement; Multispectral object detection; Pseudo-depth feature; Monocular depth estimation

[†]Corresponding author.

Contents

1 Introduction **1**

2 Method **3**
2.1 Extraction of spatial pseudo-depth features 3
2.2 Fusion of pseudo-depth and RGB features 5
2.3 Experimental architecture 5

3 Experiments **8**
3.1 Experimental settings 8
3.2 Comparison and discussion of the experiment results 9
3.3 Visualization interpretation and ablation experiments 12

4 Summary and outlook **14**

1 Introduction

Object detection serves as a core task in the field of computer vision, with extensive applications including autonomous driving, security surveillance, and scene analysis. Traditionally, these methods depend solely on RGB images to identify and locate targets, ignoring three-dimensional (3D) depth information. As a result, they only perform feature extraction and detection based on visible-light images. Over time, research has evolved into two mainstream directions: convolutional neural network (CNN)-based models and vision Transformer (ViT)-based models.

CNN-based detection frameworks are known for their efficiency in feature extraction while maintaining low computational requirements [1–4]. For example, YOLO [5] predicts multiple bounding boxes and class labels in a single forward pass of the network. Unlike anchor-based methods, CenterNet [6] utilizes keypoint detection, resulting in a simpler and more streamlined process. Vision Transformer (ViT)-based networks, on the other hand, use an encoder-decoder architecture to capture global context. It allows for better modeling of long-range dependencies, though at the expense of increased computational demands [7–13]. Among these, the DETR [14] eliminates many traditional pre- and post-processing steps by directly matching predictions to ground-truth labels using bipartite matching. Building on this, RT-DETR [15] designs a hybrid encoder to accelerate the inference speed via intra-scale and cross-scale feature interactions. It also adopts an uncertainty-minimized query selection mechanism to generate high-quality initial queries, which further improve the detection accuracy.

The detection methods discussed so far rely solely on two-dimensional (2D) RGB images for prediction. While these approaches have demonstrated promising results, they

have notable limitations. RGB images provide detailed color and texture information of the target, but they lack explicit 3D spatial details [16]. As the key information of the third dimension, depth information offers crucial spatial and temporal cues that can significantly enhance object detection and recognition [17–19]. Without this data, the perceptual capabilities of deep networks remain restricted. To overcome this challenge, recent research has incorporated depth data to supplement spatial and structural information, leading to improved detection accuracy.

Researchers have explored multi-modal fusion between RGB and depth. Specifically, DMRANet [20] utilizes residual connections to process 3D camera data, combining and fusing features from both RGB and depth streams at multiple levels. This approach enhances the complementary use of these two modalities. DETR3D [21] adapts the Transformer-based detection framework from 2D to 3D, enabling the detection of objects from multiple views in three-dimensional space. The PETR framework [22] performs direct 3D detection using multi-view images. It first divides the camera's field of view into shared grid coordinates, then encodes these grid coordinates into 3D space and integrates them with image feature information, thereby improving the model's spatial understanding ability. RadarPillars [23] introduces a pillar-based detection network, that efficiently extracts features from radar point clouds by decomposing radial velocity data, enabling high-speed object detection. Beyond depth information, many studies have demonstrated that incorporating additional features can further enhance detection performance. For example, in remote sensing applications, combining infrared and hyperspectral data has proven effective in boosting detection accuracy [24–28]. Meanwhile, multi-scale fusion for multimodal data also contributes to improving the performance of object detection[29–31].

Currently, 3D data are mainly obtained through methods such as structured light, LiDAR, and stereo vision systems [32–37]. However, these data often require additional processing, such as registration, before they can be effectively used in detection tasks. Although these multi-sensor methods effectively utilize 3D spatial features, they often face practical challenges such as high hardware costs, complex calibration procedures, limited sensor resolution, and increased technical complexity. These issues hinder their widespread deployment, especially in real-world scenarios where environmental conditions and budget constraints make multi-sensor systems impractical. To address these limitations, this work explores the possibility of incorporating depth information using only monocular RGB images. We propose a dedicated data preprocessing pipeline that generates pseudo-depth maps through monocular depth estimation model without needing additional distance sensor. This approach leverages the ability of monocular depth estimation to predict dense, reliable depth maps directly from RGB images, providing a practical and efficient foundation for enhancing object detection without relying on multiple sensors [38–40].

We design a dual-branch network architecture to evaluate the effectiveness of the proposed method. Our approach begins by generating high-quality pseudo-depth maps from RGB images using a monocular depth estimation network [41–46]. The pseudo-depth maps, alongside the original RGB images, are fed into the dual-branch architecture consisting of two identical backbone. The feature maps produced by both branches are then concatenated and fused, allowing the model to learn combined information of the optical appear-

ance and the pseudo-depth.

The proposed approach removes the requirement for additional depth sensors such as 3D cameras or LiDAR, thereby significantly lowering deployment costs and reducing technical complexity. By relying solely on monocular RGB images, our method utilizes both visual and pseudo-depth cues, offering a simple and cost-effective preprocessing pipeline that is easy to extend. Experiments on the M^3FD [47] and COCO [48] datasets show that incorporating pseudo-depth features improves detection accuracy compared to traditional RGB-based methods. Additionally, the model maintains a relatively simple architecture, making it straightforward to integrate into various network frameworks.

The main contributions of this paper are as follows:

1. We introduce a novel data preprocessing approach to enhance objection features that leverages pseudo-depth estimation from original RGB input images. The method has improved scene understanding capability and detection accuracy without additional sensors, making it low-cost and easy to implement;
2. We design a dual-branch feature fusion network that employs multi-scale concatenation and fusion strategies, which effectively reduces feature redundancies and facilitates the integration of pseudo-depth information and optical features;
3. Through extensive experiments on the M^3FD and COCO datasets, we demonstrate that incorporating pseudo-depth features can significantly improve object detection performance across multiple complex scenarios.

The rest of this work is organized as follows. Section 2 describes the proposed pseudo-depth-based data preprocessing method and the design of the feature fusion network. Section 3 details the experimental results and provides an analysis of their implications. Finally, Section 4 summarizes the conclusions of the study.

2 Method

This section introduces a data preprocessing approach that combines pseudo-depth information with monocular RGB images to enhance object detection accuracy. The section elaborates on the fusion strategy, overall framework, and key technical details of the proposed method. To evaluate its effectiveness, YOLO11n is used as the baseline model.

2.1 Extraction of spatial pseudo-depth features

Spatial depth information provides critical cues for object detection, particularly in scenarios involving occlusion, varying illumination, and cluttered backgrounds. As shown in Fig. 1, pseudo-depth maps can highlight object contours in low-light environments. Traditional methods for measuring depth include active sensors such as LiDAR and stereo camera systems. These distance sensors can deliver high-precision three-dimensional spatial data. However, the high costs, large physical sizes, and susceptibility to adverse environmental conditions have limited their widespread use in practical problems [49–52]. To address these limitations, we adopt a monocular depth estimation approach which infers

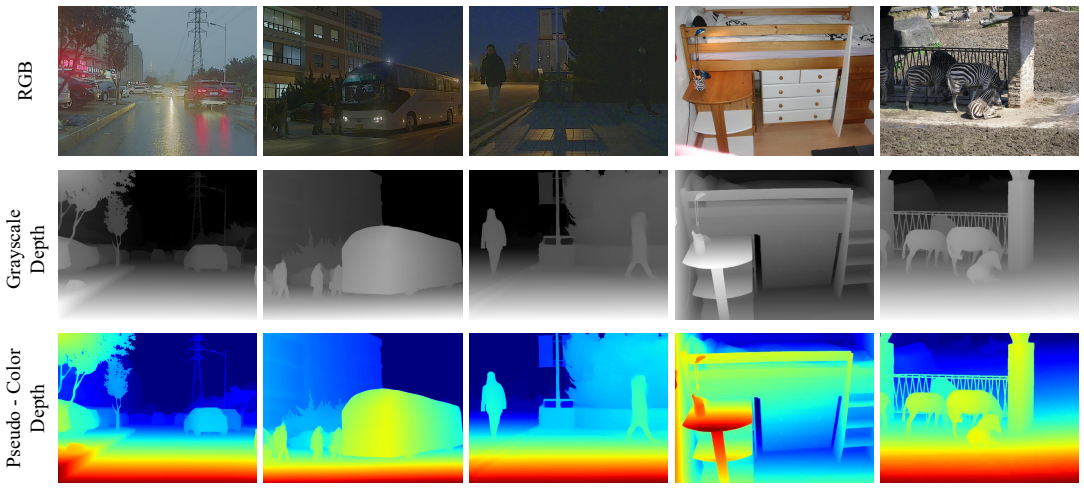

Fig. 1: Example images from the M^{3}FD (left three columns) and COCO (right two columns) dataset. From top to bottom, the rows show the original RGB images, pseudo-depth maps generated by the monocular depth estimation (MDE) model, and their pseudo-color renderings, respectively.

the pseudo-depth information from only the corresponding RGB images. It allows us to model scene geometry without additional hardware.

Here we use the Depth Anything V2 [53], an excellent monocular depth estimation model to validate our proposed method, whose network architecture is shown in Fig. 2. This model uses a two-stage training process. It first pre-trains on synthetic data to obtain the Teacher Model which then produces large-scale pseudo-annotated real images [54]. After being trained on these pseudo-annotated images, the model behaves quite well in complex environments. The model is also able to accurately detect details such as transparent objects and thin structures. Its performance on standard benchmarks is shown in Tab. I.

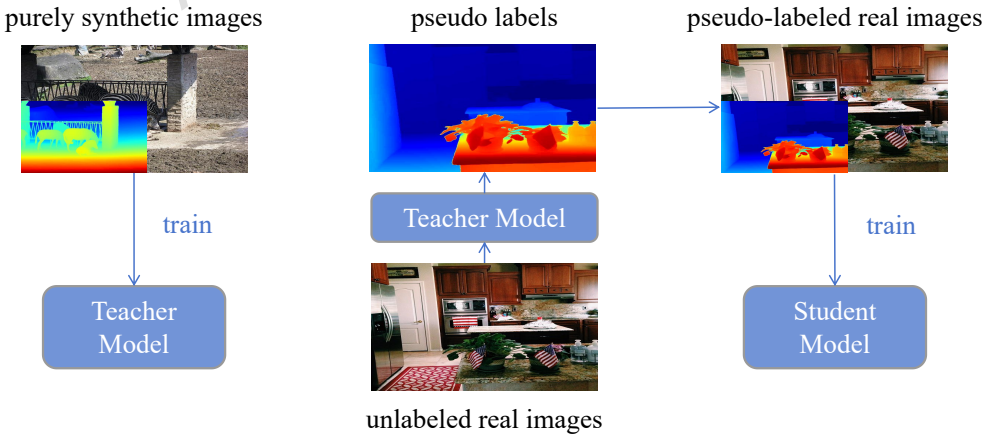

Fig. 2: The overall architecture of Depth Anything V2 [53].

The dense and continuous pseudo-depth maps produced by Depth Anything V2 contain pixel-level relative depth information. These maps complement the features of RGB images and improve scene understanding and object localization.

Tab. I: Zero-shot relative depth estimation for Depth Anything V2 [53], where the δ_1 metric representes proportion of pixel points with $\max(\frac{\hat{d}}{d}, \frac{d}{\hat{d}}) < 1.25$.

Model	KITTI	NYU-D	Sintel	ETH3D	DIODE
AbsRel	0.074	0.045	0.487	0.131	0.066
δ_1	0.946	0.979	0.752	0.865	0.952

2.2 Fusion of pseudo-depth and RGB features

To effectively leverage the complementary information from RGB images and pseudo-depth maps, this work employs a fusion strategy that utilizes parallel feature extraction networks [55, 56]. Fusion methods are typically classified into early, middle, and late fusion based on the stage at which they combine features. Early fusion techniques, such as CrossFuse [57] and DIVFusion [58], merge raw data or low-level features directly, often leading to high computational costs and limited ability to capture modality-specific details. Conversely, late fusion [59] approaches integrate high-level features or detection outputs, but they limit cross-modal interaction and collaboration.

Middle fusion offers a balance by integrating semantic information at intermediate feature levels. This approach preserves modality-specific characteristics while enabling effective cross-modal interaction [60–63]. For example, Transformer-guided cross-modal fusion (CMF) [64] can learn long-range dependencies, supporting both intra-modal and inter-modal feature integration. Similarly, MBNet [65] employs modality-aware modules and feature alignment to enhance complementarity and reduce discrepancies during fusion. Considering these advantages, this work adopts a middle fusion strategy. The overall framework is shown in Fig. 3, where the feature maps of the RGB branch f_{rgb} and the pseudo-depth one f_{depth} are concatenated to produce the fused feature map f.

2.3 Experimental architecture

The foundation of the system is built on the YOLO11n detection framework [66] for feature fusion. In YOLO11, The backbone network extracts features at multiple scales, capturing both semantic and spatial information. It consists of convolutional and fusion modules, incorporating multi-scale pooling and attention mechanisms. This backbone produces three key feature maps, labeled $f^{(3)}$, $f^{(4)}$, and $f^{(5)}$, each with different strides. In the neck network, bidirectional feature fusion combines high-level and low-level spatial features. This facilitates effective multi-scale object detection. The network outputs three fused feature maps, $\hat{f}^{(3)}$, $\hat{f}^{(4)}$, and $\hat{f}^{(5)}$. Finally, the detection head converts these fused maps into detection results, providing the final predictions.

The input to the proposed framework includes both RGB images and pseudo-depth maps generated from the original RGB ones through monocular depth estimation model.

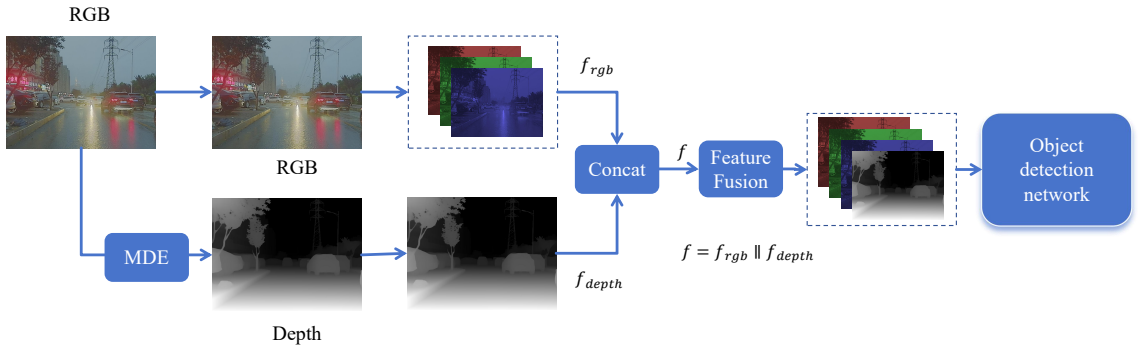

Fig. 3: The overall structure of the experimental network is as follows. An RGB image is used to generate a pseudo-depth map via the MDE, which is then fed into the network through a dual-branch structure for feature fusion.

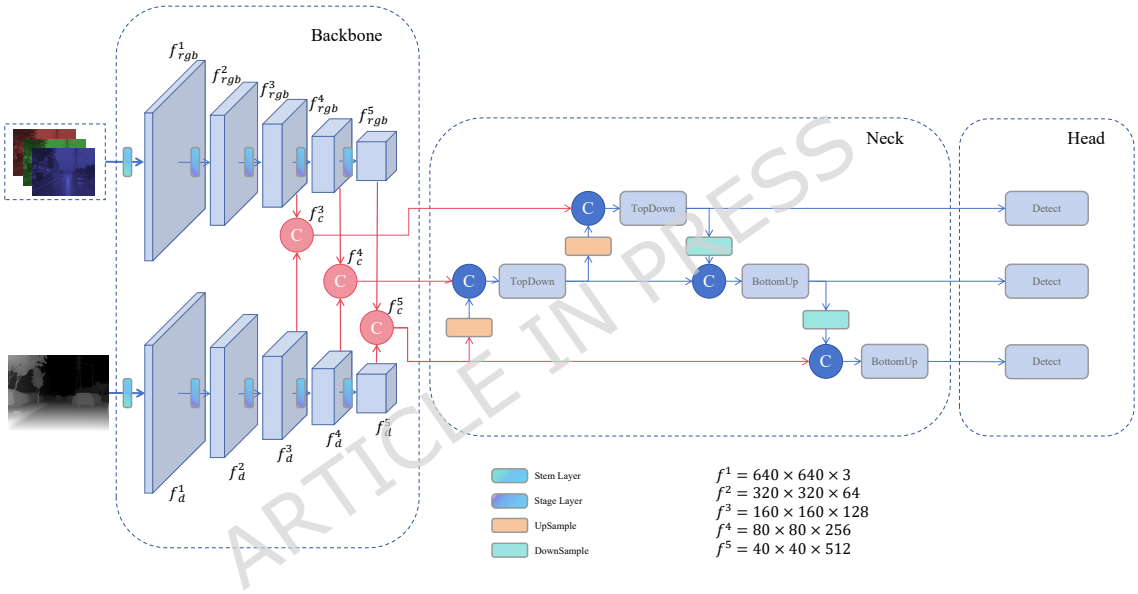

Fig. 4: The network structure. The RGB and pseudo-depth features are concatenated at the $f^{(3)}$, $f^{(4)}$, and $f^{(5)}$ layer. Then the features are fused at multiple scales in the neck network. The three produced detection-scale features are finally used as input to the detection module.

To enable effective feature fusion, the framework incorporates two core modules, namely, a dual-branch feature extraction module and a mid-level feature fusion module. The overall architecture is shown in Fig. 4.

The experimental framework features a dual-branch structure that processes RGB and pseudo-depth images separately. Each input is passed through an independent backbone network, designed similarly to the original YOLO11 backbone, to extract multi-layer features. The feature maps of layers $f^{(3)}$, $f^{(4)}$, and $f^{(5)}$ are concatenated to form fused feature

maps, labeled as $f_c^{(3)}$, $f_c^{(4)}$, and $f_c^{(5)}$. The concatenation operation is formally defined as

$$f_c^{(n)} = f_{\text{rgb}}^{(n)} \parallel f_{\text{depth}}^{(n)}, \tag{2.1}$$

where $n = 3, 4, 5$ and f_n denotes the n-th feature layer.

These fused features undergo multi-scale processing in the detection head, which follows the same structure as the original YOLO11 head. For example, the output of the $f_c^{(3)}$ layer after passing through the detection head is denoted as $F_{f^{(3)}}^{\text{head}}$. The fusion formulation for $F_{f^{(3)}}^{\text{head}}$ is expressed as

$$F_{f^{(4)}}^{*} = f_c^{(4)} \oplus g^{\text{Upsample}}(f_c^{(5)}, 2), \tag{2.2}$$

$$F_{f^{(3)}}^{*} = f_c^{(3)} \oplus g^{\text{Upsample}}\left(g^{\text{fusion}}(F_{f^{(4)}}^{*}, 512), 2\right), \tag{2.3}$$

$$F_{f^{(3)}}^{\text{head}} = g^{\text{fusion}}(F_{f^{(3)}}^{*}, 256), \tag{2.4}$$

where $g^{\text{Upsample}}(f, s)$ denotes upsampling f by a scale factor of s; $g^{\text{fusion}}(f, s)$ denotes compressing f to the target dimension s using the fusion module in YOLO11; F^{head} represents the output of features at the corresponding scale in the head. The fused features at three scales are then fed into YOLO11's detection module to produce the final detection results. Meanwhile, we performed similar modifications on YOLOv8 and RT-DETR by fusing their $f^{(3)}$, $f^{(4)}$, and $f^{(5)}$ feature layers. As illustrated in Fig. 4, the input feature dimensions of the five feature layers are annotated in the bottom-right corner of the figure. The red-highlighted sections (backbone and neck) represent operations that differ from the standard YOLO11n design in our setup. The remaining components, shown in blue, maintain the same structure as YOLO11n.

The loss function in YOLO11 comprises three components [66, 67]: the object classification loss (Loss_{cls}), the localization loss (Loss_{loc}), and the distribution focal loss (Loss_{dfl}). The classification loss is used to control the correct class label for each object. It is computed by using the Binary Cross-Entropy (BCE) [68]. The localization loss evaluates the difference between the predicted bounding box and the actual bounding box. The localization loss consists of two parts, the center point coordinate loss and the width-height parameter loss. The former one evaluates the squared Euclidean distance between the predicted and ground-truth center coordinates of the bounding box,

$$\text{Loss}_{\text{center}} = \sum (\tilde{x}_i - x_i)^2 + (\tilde{y}_i - y_i)^2, \tag{2.5}$$

while the latter one evaluates the difference between the predicted and the actual width-height parameters,

$$\text{Loss}_{w-h} = \sum \left(\sqrt{w_i} - \sqrt{\tilde{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\tilde{h}_i}\right)^2, \tag{2.6}$$

where the tilde variables denote the predicted ones. The distribution focal loss is introduced to address the issue of class imbalance in object detection and to improve the performance

in dealing with small targets and difficult samples [69]. Formally, Loss_{dfl} is defined as

$$\text{Loss}_{\text{dfl}}(P_i, P_{i+1}) = -\left[(y_{i+1} - y)\log(P_i) + (y - y_i)\log(P_{i+1})\right], \tag{2.7}$$

where

$$P_i = \frac{y_{i+1} - y}{y_{i+1} - y_i}, P_{i+1} = \frac{y - y_i}{y_{i+1} - y_i}, \tag{2.8}$$

where i denotes the index of the discrete coordinate points; y denotes the ground-truth location value; y_i and y_{i+1} are two consecutive discrete points closest to y. The total loss is computed as a weighted sum of the three components:

$$\text{Loss}_{\text{total}} = \lambda_{\text{cls}}\text{Loss}_{\text{cls}} + \lambda_{\text{loc}}\text{Loss}_{\text{loc}} + \lambda_{\text{dfl}}\text{Loss}_{\text{dfl}}, \tag{2.9}$$

where λ_{cls}, λ_{loc}, and λ_{dfl} are balance weights that control the contributions of the classification loss, localization loss, and distribution focal loss, respectively. These three weight parameters are set to the standard values used in the YOLO11 framework.

The combined loss function provides multi-level supervision. It helps reduce class imbalance and improves detection accuracy, especially in complex environments.

3 Experiments

We conducted experiments on COCO and M^3FD datasets to validate the effectiveness of our proposed method. We compared the performance of the RGB-only detection and our pseudo-depth fusion methods. The results show that the proposed scheme can significantly improve the detection performance on both datasets. The proposed method also remains stable across different training iterations.

3.1 Experimental settings

In this work, we evaluate the effectiveness of the proposed algorithm using the datasets M^3FD and COCO. The M^3FD data is a multi-modal dataset comprising synchronized thermal infrared and visible light RGB images captured via binocular optical systems and infrared sensors. This dataset emphasizes multi-modal features across various complex scenarios and different pixel variations. The images are resolved at 1024×768 pixels. Annotations cover six categories, 'Person', 'Car', 'Bus', 'Motorcycle', 'Lamp', 'Truck'. The dataset is split such that 80% of the images are randomly selected for training and the remaining 20% for testing. COCO is a large dataset for object detection. It contains images from natural and complex daily scenes. COCO dataset includes 328,000 images and 2.5 million annotated instances. The annotations cover 80 object categories, such as 'Car', 'Bottle', and 'Cat'. For the experiments, 20,000 images were randomly sampled as the training set and 2,000 ones were used as the test set.

To quantitatively evaluate the performance of the proposed method, here we employ the detection metrics F_1-score, mAP_{50} (mean Average Precision at IoU=0.50) and

$\text{mAP}_{50:95}$ (mean Average Precision over IoU thresholds from 0.50 to 0.95) [70]. The F_1-score is defined as

$$F_1 = \frac{2PR}{P+R}, \tag{3.1}$$

which comprehensively considers both precision (P) and recall (R) with the same weight. In our experiments, the F_1-score is denotes the mean F_1-score across all categories. IoU measures the degree of overlap between a single predicted bounding box and a single ground-truth box. For the prediction box A and ground truth box B, the IoU is defined as

$$\text{IoU} = \frac{A \cap B}{A \cup B}. \tag{3.2}$$

The metric Average Precision (AP) denotes the area under the precision-recall curve for each class. In practice AP is the precision averaged across all recall values between 0 and 1. For the ith category,

$$\text{AP}_i = \int_0^1 P_i(R_i)\,\mathrm{d}R_i = \sum_{k=0}^{n} P_i(k)\Delta R_i(k), \tag{3.3}$$

and the Mean Average Precision (mAP) denotes the mean of AP across all C classes,

$$\text{mAP} = \frac{1}{C}\sum_{i=1}^{C} \text{AP}_i. \tag{3.4}$$

The experiments are performed on a workstation equipped with an NVIDIA Tesla P40 GPU, utilizing the PyTorch 1.12 framework. All input images are resized to 640×640 pixels using the letterbox method [71]. The model training employs the Stochastic Gradient Descent (SGD) [72] optimizer with a batch size of 16. The first 3 epochs serve as warm-up with an initial learning rate of 0.005, followed by linear decay to a minimum of 5×10^{-5}. Data augmentation techniques such as Mosaic [73] and MixUp [74] are employed to improve generalization and robustness.

3.2 Comparison and discussion of the experiment results

This study uses YOLO11n as the baseline model to assess the effectiveness of the proposed approach. For comparison, the classic YOLOv8-n and RT-DETR models are also employed to evaluate the generalization capabilities of the proposed method. All training procedures are conducted from scratch with no pre-trained weights used. We compare the performance of the model when using only RGB images against the performance when fusing pseudo-depth information. The results, obtained on COCO and M^3FD datasets, are summarized in Tab. II.

The results indicate that incorporating pseudo-depth features significantly enhances the model's detection performance on both the M^3FD and COCO datasets. Specifically, in the YOLO11n based experiments, mAP_{50} metric can improve 3.8 and 8.0 percentage points

Tab. II: Comparison of our proposed scheme on the YOLO11n, the YOLOv8-n and the RT-DETR detection models on the COCO and the M^{3}FD data.

Data	Model	Method	Paras(M)	Flops(G)	mAP$_{50}$(%)	mAP$_{50:95}$(%)	F_1(%)
M^{3}FD	YOLO11n	RGB	2.6	6.3	73.7	47.8	73.1
		RGB-pD	3.8	9.3	(+3.8)77.5	(+3.2)51.0	75.7
	YOLOv8-n	RGB	3.0	8.1	77.4	50.7	86.5
		RGB-pD	4.4	11.3	(+2.5)79.9	(+2.3)53.0	78.7
	RT-DETR	RGB	41.9	125.7	82.5	53.9	79.3
		RGB-pD	66.3	194.0	(+1.3)83.8	(+1.5)55.4	81.8
COCO	YOLO11n	RGB	2.6	6.5	41.8	28.6	43.3
		RGB-pD	3.8	9.5	(+8.0)49.8	(+7.1)35.7	50.9
	YOLOv8-n	RGB	3.2	8.8	41.3	28.0	42.5
		RGB-pD	4.5	12.0	(+8.0)49.3	(+7.2)35.2	50.1
	RT-DETR	RGB	42.1	126.0	37.7	24.6	42.5
		RGB-pD	66.5	194.4	(+8.5)46.2	(+7.8)32.4	49.3

on the M^{3}FD and COCO dataset, respectively. As shown in the table, similar performance gains are also observed in experiments using YOLOv8 and RT-DETR frameworks.

Tab. III and Tab. IV show the category-specific detection results for the YOLO11n-based experiments on M^{3}FD and COCO datasets, respectively. The rows present AP$_{50}$ and AP$_{50:95}$ scores for each category on the M^{3}FD dataset; while six randomly chosen categories are shown for COCO dataset. Across all categories, the model incorporating pseudo-depth can consistently outperform the original one.

Tab. III: Category-specific detection results on M^{3}FD.

Class		People	Car	Lamp	Bus	Motorcycle	Truck
AP$_{50}$(%)	RGB	64.1	88.1	85.9	71.8	56.7	75.7
	RGB-pD	67.5	89.9	87.4	74.0	63.7	82.5
AP$_{50:95}$(%)	RGB	33.9	63.6	68.1	36.1	33.2	52.2
	RGB-pD	36.2	65.4	71.9	39.0	35.7	58.1

Tab. IV: Category-specific detection results on COCO.

Class		Person	Airplane	Motorcycle	Fire Hydrant	Bear	Toilet
AP$_{50}$(%)	RGB	68.2	71.3	47.6	61.7	62.3	61.2
	RGB-pD	71.8	80.3	59.8	69.7	86.9	72.6
AP$_{50:95}$(%)	RGB	44.8	52.5	29.5	50.0	51.3	50.9
	RGB-pD	49.3	57.2	36.8	59.5	65.9	60.8

In the first two rows of Fig. 5, we show the predicted results without and with fusing the pseudo-depth feature. Integrating pseudo-depth information helps the model define

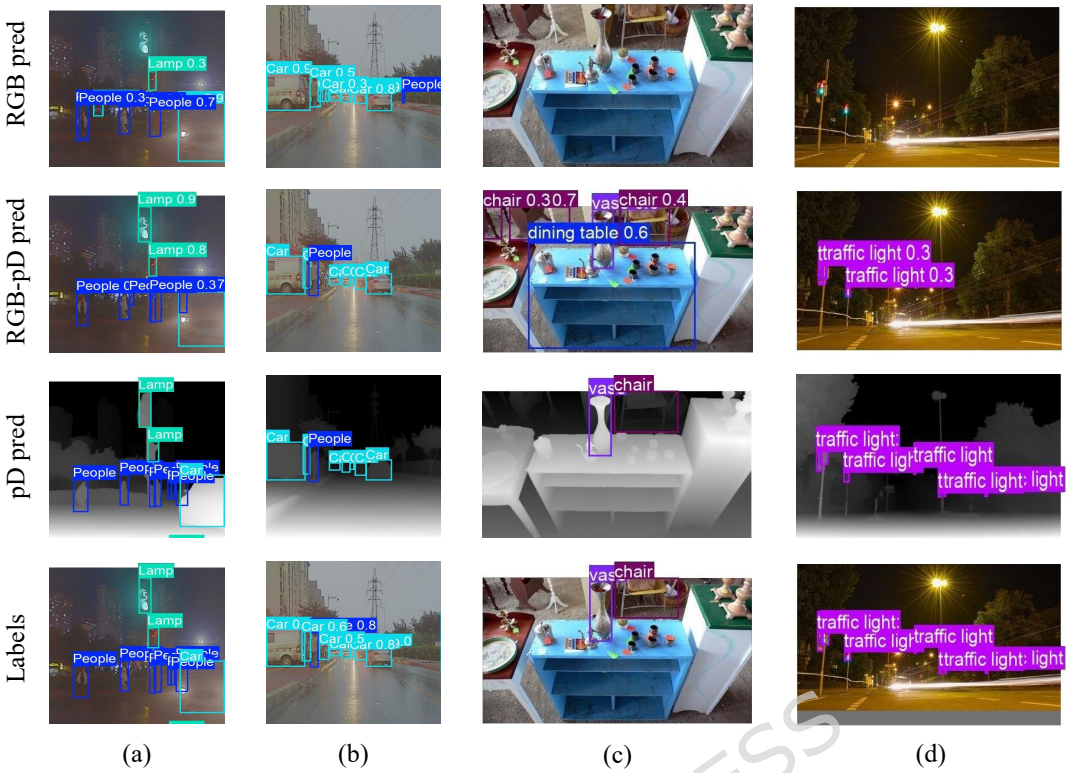

Fig. 5: Visualization of the experimental results. The first two rows show the predicted results by the RGB-only features and the RGB fused with pseudo-depth features. The last two rows present the ground-truth annotations on the pseudo-depth images and the RGB images, respectively. The columns (a) and (b) show results on the M^3FD dataset, whereas (c) and (d) show results on the COCO data.

object boundaries more clearly. It also reduces background noise, which decreases missed detections. This is especially noticeable in low-light or cluttered scenes. For example, in the M^3FD sample (Fig. 5 a), the RGB-only model missed a lamp in the upper middle part of the image because of low lighting at night. However, the pseudo-depth map clearly shows the lamp's contours. The detector that uses pseudo-depth can accurately identify the target, even when it blends into the background. This improvement comes from the pseudo-depth map's ability to enhance object edges. Depth-guided cues help distinguish targets from noisy backgrounds, thus improving localization and detection reliability. It is important to note that adding pseudo-depth processing increases the computational load by approximately 46%. Despite this, the significant performance gains show that this trade-off is acceptable for practical applications.

Fig. 6 shows the changes in mAP_{50} and $mAP_{50:95}$ during training on YOLO11n based model. Even at the early stages, incorporating pseudo-depth information significantly improves the detection performance. This benefit continues as the training progresses. The results demonstrate that the proposed method is effective and can produce performance gains even with limited training time. The performance difference between the two datasets

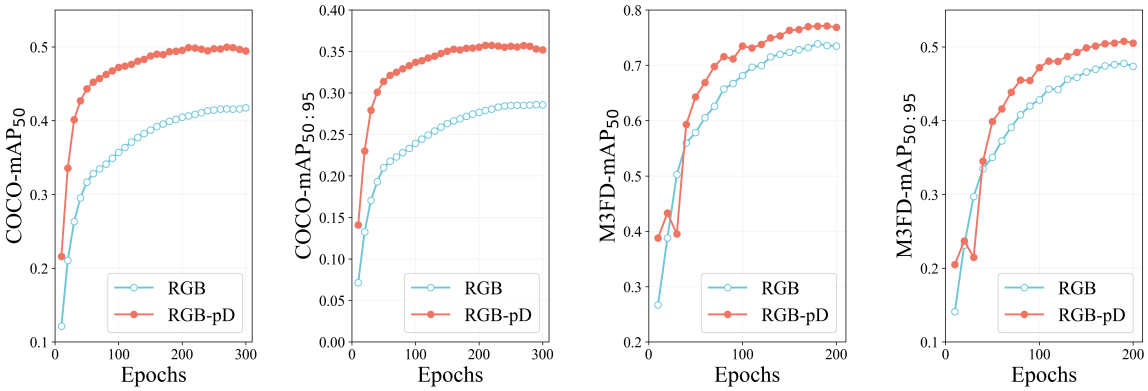

Fig. 6: The metrics mAP_{50} and $mAP_{50:90}$ versus the training epochs on the datasets M^3FD and COCO. The red solid (blue hollow) circle denotes the results based on the RGB-pD (RGB-only) features.

is likely related to their scene characteristics. The COCO dataset contains many indoor and close-range images. In these scenes, depth estimation tends to be more accurate. This leads to a higher quality of pseudo-depth features and larger gains in detection performance. In contrast, the M^3FD dataset includes more outdoor scenes. Factors such as changing lighting conditions and long-distance targets make accurate depth estimation more difficult. As a result, the overall benefit of pseudo-depth fusion is limited.

Tab. V: Inference Efficiency of Different Models.

Model	Method	Latency(s, $\times 10^{-4}$)	$FPS_{bs=1}$
YOLO11n	RGB	102.5 ± 16.0	97.6
	RGB-pD	152.4 ± 27.1	65.6
YOLOv8-n	RGB	75.0 ± 12.6	133.4
	RGB-pD	104.3 ± 3.8	95.8
RT-DETR	RGB	262.1 ± 15.5	38.2
	RGB-pD	351.3 ± 43.4	28.5

Tab. V shows the inference efficiency of several models. All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU, with a batch size of 1, and the results are calculated based on 1000 test runs. The table indicates that the model with pseudo-depth integration experiences a reduction in inference efficiency compared to the original model. This decline is primarily due to the increased number of parameters resulting from the added branch, reflecting a trade-off between enhanced feature representation and computational speed.

3.3 Visualization interpretation and ablation experiments

GradCAM is employed to visualize the prediction effects of the network. Fig. 7 illustrates the GradCAM visualization results. Specifically, the first column shows the original images

with manual bounding boxes, and the second column presents the corresponding pseudo-depth maps of the original images. The third and fourth columns display the visualization heatmaps of the detection heads for the original YOLO11 and the YOLO11 integrated with pseudo-depth information, respectively. The last column shows the visualization heatmaps of the $f_c^{(3)}$, $f_c^{(4)}$, and $f_c^{(5)}$ feature layers. It can be observed that after introducing the depth maps, the attention of the fusion layer is more concentrated on the boundaries of objects with obvious foreground-background relationships. Meanwhile, the attention of the detection head is accordingly focused on these regions, which improves the performance of object detection to a certain extent.

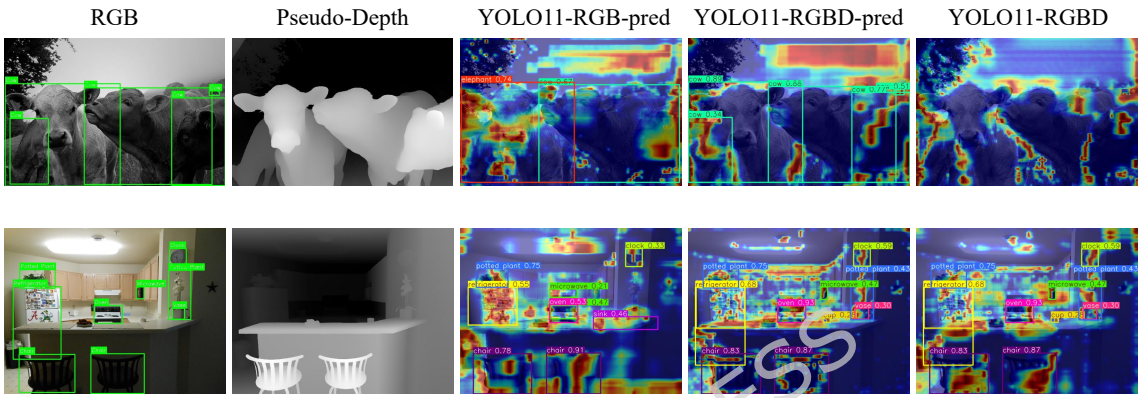

Fig. 7: The GradCAM visualization of the two models on COCO dataset.

Since the comparison of RGB and RGB-pD already behaves like an ablation study, we design the ablation study by replacing the pseudo-depth maps by grayscale maps derived from the original RGB images to evaluate the effectiveness of the proposed module. Tab. VI shows the ablation experimental results based on the YOLO11 framework.

Tab. VI: Ablation experiments of YOLO11 based framework on M^{3}FD and COCO datasets.

Dataset	Method	mAP$_{50}$(%)	mAP$_{50:95}$(%)
M^{3}FD	RGB	73.7	47.8
	RGB-pD	(+3.8)77.5	(+3.2)51.0
	RGB-gray	(+3.6)77.3	(+3.1)50.9
COCO	RGB	41.8	28.6
	RGB-pD	(+8)49.8	(+7.1)35.7
	RGB-gray	(+1.2)43.0	(+1.0)29.6

Results on COCO dataset confirm the validity of this ablation study. Compared with the RGB-gray model, the RGB-pD model achieves a 6.8 improvement in mAP$_{50}$, which indicates that introducing pseudo-depth information significantly boosts detection performance. On M^{3}FD dataset, the ablation study reveals only a modest performance difference

between the proposed RGB-pD model and the RGB-gray model. This similarity in results may be due to the nature of the M^{3}FD dataset. The M^{3}FD dataset primarily features road scenes with relatively uniform but distant object characteristics. In these conditions, the obtained gray maps are much more similar with the pseudo-depth ones. The additional pseudo-depth information thus provides limited complementary benefit, resulting in a minor performance gap. The improvements on the COCO dataset demonstrate that the proposed module can substantially enhance detection performance in more diverse and complex data scenarios, providing strong evidence of its value for model optimization.

Meanwhile, we also observe that converting RGB images to grayscale before inputting them into the network leads to a noticeable improvement in detection performance. This improvement likely arises because grayscale input reduces the model's reliance on color information, encouraging it to instead learn structural features such as edges and textures. Such features generally exhibit stronger generalization ability than color cues.

The experimental results confirm that the proposed pseudo-depth fusion method significantly enhances the performance of object detection models, particularly in scenarios where accurate depth estimation can be achieved. Moreover, the integration of pseudo-depth features offers several distinct advantages:

1. Enhance the perception of object boundaries and contours, reduce missed detections, and improves bounding box regression accuracy;
2. Boost the model robustness under challenging conditions such as low illumination and object occlusion;
3. Provide a low-cost, low-complexity scheme for integrating depth information, which is easy to implement and deploy.

4 Summary and outlook

This work introduces a data preprocessing approach to improve RGB-based visual object detection based on estimated depth information. Unlike costly and complex real depth sensors, the method generates pseudo-depth maps from original RGB input images. We employ a dual-branch feature fusion strategy to concurrently extract and combine high-level features from both RGB and pseudo-depth inputs. Several comparative experimental results on the COCO and the M^{3}FD datasets confirm that the proposed scheme can consistently improve detection accuracy and robustness across multiple scenarios. The mAP metric improvement can reach 8 percentage points on COCO dataset. The advantages of the proposed scheme are twofold: (I) it provides a simple and low-cost solution without extra detection sensors; (II) it can be easily embedded into most machine learning models to definitely improve the detection performance.

On the hand, the results give new hint to further improve the visual performance besides the traditional updating in model architecture and data engineering. It reveals that though the deep neutral network can spontaneously extract complex abstract information from the original input, it can bring more benefit to feed the distilled features directly into the network for specific tasks. More independent features are supposed to improve

the performance in machine learning tasks. Besides the passive RGB spectral features and the depth feature, the infrared information can behave as an independent active spectral feature to make positive effects. The (pseudo-)infrared features can be obtained similarly as the (pseudo-)depth one based on the deep neural network.

Acknowledgments

This work is supported by the National Key R&D Program of China (2022YFA1604803), the National Major in High Resolution Earth Observation (68-Y50G07-9001-22/23), and the Natural Science Basic Research Program of Shaanxi (2025JC-YBMS-020).

Data availability

The datasets generated and/or analysed during the current study are available in the GitHub repository with link https://github.com/htyb275/Pseudo-Depth-Detection.

References

[1] R. Girshick, J. Donahue, T. Darrell and J. Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation*, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014, DOI.

[2] S. Ren, K. He, R. Girshick and J. Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39** (2017) 1137.

[3] J. Dai, Y. Li, K. He and J. Sun, *R-fcn: object detection via region-based fully convolutional networks*, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, (Red Hook, NY, USA), p. 379–387, Curran Associates Inc., 2016.

[4] H. Law and J. Deng, *Cornernet: Detecting objects as paired keypoints*, in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss, eds., (Cham), pp. 765–781, Springer International Publishing, 2018.

[5] P. Jiang, D. Ergu, F. Liu, Y. Cai and B. Ma, *A review of yolo algorithm developments*, *Procedia Computer Science* **199** (2022) 1066.

[6] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang and Q. Tian, *Centernet: Keypoint triplets for object detection*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October, 2019.

[7] X. Zhu, W. Su, L. Lu, B. Li, X. Wang and J. Dai, *Deformable detr: Deformable transformers for end-to-end object detection*, 2010.04159.

[8] X. Tong, W. Feng, W. Xu, C.-H. Chang, G.-L. Wang and Q. Li, *Meson Properties and Symmetry Emergence Based on the Deep Neural Network*, *Chin. Phys. Lett.* **43** (2026) 020201 [2509.17093].

[9] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski et al., *Emerging properties in self-supervised vision transformers*, in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

[10] Z. Niu, G. Zhong and H. Yu, *A review on the attention mechanism of deep learning*, *Neurocomputing* **452** (2021) 48.

[11] Y. Chen, L. Chen, R. Xia, K. Yang and K. Zou, *Caat: Image super-resolution algorithm via channel attention and transformer*, *Array* **28** (2025) 100628.

[12] S. Meerits, D. Thomas, V. Nozick and H. Saito, *Fusionmls: Highly dynamic 3d reconstruction with consumergrade rgb-d cameras*, *Computational Visual Media* **4** (2018) 287.

[13] X. Chu, J. Deng, J. Ji, Y. Zhang, H. Li and Y. Zhang, *Oa-det3d: Embedding object awareness as a general plug-in for multi-camera 3d object detection: X. chu et al.*, *International Journal of Computer Vision* **133** (2025) 8022.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, *End-to-end object detection with transformers*, in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox and J.-M. Frahm, eds., (Cham), pp. 213–229, Springer International Publishing, 2020.

[15] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang et al., *Detrs beat yolos on real-time object detection*, in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974, 2024, DOI.

[16] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang and S. Kwong, *Going from rgb to rgbd saliency: A depth-guided transformation model*, *IEEE Transactions on Cybernetics* **50** (2020) 3627.

[17] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang et al., *Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection*, *IEEE Transactions on Image Processing* **31** (2022) 6800.

[18] Y. Piao, Z. Rong, M. Zhang, W. Ren and H. Lu, *A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection*, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, p. 9057 – 9066, 2020, DOI.

[19] T. Wang, X. Zhu, J. Pang and D. Lin, *Fcos3d: Fully convolutional one-stage monocular 3d object detection*, in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 913–922, 2021.

[20] Y. Piao, W. Ji, J. Li, M. Zhang and H. Lu, *Depth-induced multi-scale recurrent attention network for saliency detection*, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7253–7262, 2019, DOI.

[21] Y. Wang, V.C. Guizilini, T. Zhang, Y. Wang, H. Zhao and J. Solomon, *Detr3d: 3d object detection from multi-view images via 3d-to-2d queries*, in *Proceedings of the 5th Conference on Robot Learning*, A. Faust, D. Hsu and G. Neumann, eds., vol. 164 of *Proceedings of Machine Learning Research*, pp. 180–191, PMLR, 08–11 Nov, 2022, https://proceedings.mlr.press/v164/wang22b.html.

[22] Y. Liu, T. Wang, X. Zhang and J. Sun, *Petr: Position embedding transformation for multi-view 3d object detection*, in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G.M. Farinella and T. Hassner, eds., (Cham), pp. 531–548, Springer Nature Switzerland, 2022.

[23] A. Musiat, L. Reichardt, M. Schulze and O. Wasenmüller, *Radarpillars: Efficient object detection from 4d radar point clouds*, in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1656–1663, IEEE, 2024.

[24] H. Song, J. Xie, Y. Duan, X. Xie, Y. Zhou and W. Wang, *Cmkd-net: a cross-modal knowledge distillation method for remote sensing image classification*, *Advances in Space Research* **75** (2025) 8515.

[25] W. Zhou, Y. Cai, X. Dong, F. Qiang and W. Qiu, *Adrnet-s*: Asymmetric depth registration network via contrastive knowledge distillation for rgb-d mirror segmentation*, *Information Fusion* **108** (2024) 102392.

[26] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao et al., *Calibrated rgb-d salient object detection*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9471–9481, June, 2021.

[27] H. Song, J. Xie, L. Liang, Y. Su, Y. Xiao, X. Zhang et al., *Symmetrical learning and transferring: Efficient knowledge distillation for remote sensing image classification*, *Symmetry* **17** (2025) 1002.

[28] C.R. Qi, O. Litany, K. He and L.J. Guibas, *Deep hough voting for 3d object detection in point clouds*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October, 2019.

[29] Y. Chen, R. Xia, K. Yang and K. Zou, *Dual degradation image inpainting method via adaptive feature fusion and u-net network*, *Applied Soft Computing* **174** (2025) 113010.

[30] J. Zhang, J. Yang, Y. Qin, Z. Xiao and J. Wang, *Mgnet: Rgbt tracking via cross-modality cross-region mutual guidance*, *Neural Networks* **190** (2025) 107707.

[31] J. Zhang, S. Zhang, D. Li, J. Wang and J. Wang, *Crack segmentation network via difference convolution-based encoder and hybrid cnn-mamba multi-scale attention*, *Pattern Recognition* **167** (2025) 111723.

[32] S. Shi, X. Wang and H. Li, *Pointrcnn: 3d object proposal generation and detection from point cloud*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2019.

[33] H. Zhang, H. Jiang, Q. Yao, Y. Sun, R. Zhang, H. Zhao et al., *Detect anything 3d in the wild*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5048–5059, October, 2025.

[34] L. Yang, T. Tang, J. Li, K. Yuan, K. Wu, P. Chen et al., *Bevheight++: Toward robust visual centric 3d object detection*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025) .

[35] H. Zhang, Z. Yang, Y. Sun, L. Chen, F. Xia, F. Güney et al., *Test-time correction: An online 3d detection system via visual prompting*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **48** (2026) 3666.

[36] L. Yang, X. Zhang, J. Li, L. Wang, C. Zhang, L. Ju et al., *Sgv3d: Toward scenario generalization for vision-based roadside 3d object detection*, *IEEE Transactions on Intelligent Transportation Systems* (2025) .

[37] M. Simony, S. Milzy, K. Amendey and H.-M. Gross, *Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds*, in *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[38] X. Weng and K. Kitani, *Monocular 3d object detection with pseudo-lidar point cloud*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct, 2019.

[39] Z. Chen, Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang et al., *Graph-detr4d: Spatio-temporal graph modeling for multi-view 3d object detection*, *IEEE Transactions on Image Processing* **33** (2024) 4488.

[40] Q. Xie, Y.-K. Lai, J. Wu, Z. Wang, Y. Zhang, K. Xu et al., *Mlcvnet: Multi-level context votenet for 3d object detection*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2020.

[41] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler and V. Koltun, *Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44** (2022) 1623.

[42] J. Wang, C. Lin, L. Sun, R. Liu, L. Nie, M. Li et al., *From editor to dense geometry estimator*, 2509.04338.

[43] M. Oquab, T. Darcet, T. Moutakanni, H.V. Vo, M. Szafraniec, V. Khalidov et al., *Dinov2: Learning robust visual features without supervision*, *Transactions on Machine Learning Research* **2024** (2024) .

[44] Y. Zhou and O. Tuzel, *Voxelnet: End-to-end learning for point cloud based 3d object detection*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4490–4499, 2018.

[45] M. Liang, B. Yang, S. Wang and R. Urtasun, *Deep continuous fusion for multi-sensor 3d object detection*, in *Proceedings of the European conference on computer vision (ECCV)*, pp. 641–656, 2018.

[46] Z. Yang, Y. Sun, S. Liu, X. Shen and J. Jia, *Std: Sparse-to-dense 3d object detector for point cloud*, in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1951–1960, 2019.

[47] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong et al., *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*, pp. 5802–5811, 2022.

[48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan et al., *Microsoft coco: Common objects in context*, in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars, eds., (Cham), pp. 740–755, Springer International Publishing, 2014.

[49] D. Park, R. Ambrus, V.C. Guizilini, J. Li and A. Gaidon, *Is pseudo-lidar needed for monocular 3d object detection?*, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021) 3122.

[50] X. Ma, S. Liu, Z. Xia, H. Zhang, X. Zeng and W. Ouyang, *Rethinking pseudo-lidar representation*, in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox and J.-M. Frahm, eds., (Cham), pp. 311–327, Springer International Publishing, 2020.

[51] Z. Liu, Y. Tan, Q. He and Y. Xiao, *Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection*, *IEEE Transactions on Circuits and Systems for Video Technology* **32** (2022) 4486.

[52] H. Zhang, J.Y. Koh, J. Baldridge, H. Lee and Y. Yang, *Cross-modal contrastive learning for text-to-image generation*, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021.

[53] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng et al., *Depth anything v2*, in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak et al., eds., vol. 37, pp. 21875–21911, Curran Associates, Inc., 2024, DOI.

[54] Y. Tian, L. Fan, K. Chen, D. Katabi, D. Krishnan and P. Isola, *Learning vision from models rivals learning vision from data*, *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023) 15887.

[55] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu et al., *Transfusion: Robust lidar-camera fusion for 3d object detection with transformers*, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1090–1099, 2022.

[56] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu et al., *Bevformer: Learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47** (2025) 2020.

[57] H. Li and X.-J. Wu, *Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach*, *Information Fusion* **103** (2024) 102147.

[58] L. Tang, X. Xiang, H. Zhang, M. Gong and J. Ma, *Divfusion: Darkness-free infrared and visible image fusion*, *Information Fusion* **91** (2023) 477.

[59] H. Chen, Y. Li and D. Su, *Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection*, *Pattern Recognition* **86** (2019) 376.

[60] S.A. Deevi, C. Lee, L. Gan, S. Nagesh, G. Pandey and S.-J. Chung, *Rgb-x object detection via scene-specific fusion modules*, in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7351–7360, 2024, DOI.

[61] X. Guo, W. Zhou and T. Liu, *Contrastive learning-based knowledge distillation for rgb-thermal urban scene semantic segmentation*, *Knowledge-Based Systems* **292** (2024) 111588.

[62] J. Ma, W. Jiang, X. Tang, X. Zhang, F. Liu and L. Jiao, *Multiscale sparse cross-attention network for remote sensing scene classification*, *IEEE Transactions on Geoscience and Remote Sensing* **63** (2025) 1.

[63] D. Wan, R. Lu, Y. Fang, X. Lang, S. Shu, J. Chen et al., *Yolov11-rgbt: Towards a comprehensive single-stage multispectral object detection framework*, 2506.14696.

[64] F. Qingyun, H. Dapeng and W. Zhaokui, *Cross-modality fusion transformer for multispectral object detection*, 2111.00273.

[65] K. Zhou, L. Chen and X. Cao, *Improving multispectral pedestrian detection by addressing modality imbalance problems*, in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox and J.-M. Frahm, eds., (Cham), pp. 787–803, Springer International Publishing, 2020.

[66] G. Jocher and J. Qiu, *Ultralytics yolo11*, 2024.

[67] R. Khanam and M. Hussain, *Yolov11: An overview of the key architectural enhancements*, 2410.17725.

[68] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, *Focal loss for dense object detection*, in *2017 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV)*, IEEE International Conference on Computer Vision, pp. 2999–3007, IEEE; IEEE Comp Soc, 2017, DOI.

[69] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li et al., *Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection*, in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin, eds., vol. 33, pp. 21002–21012, Curran Associates, Inc., 2020, https://proceedings.neurips.cc/paper_files/paper/2020/file/f0bda020d2470f2e74990a07a607ebd9-Paper.pdf.

[70] J. Cartucho, R. Ventura and M. Veloso, *Robust object recognition through symbiotic deep learning in mobile robots*, in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2336–2341, 2018.

[71] J. Redmon and A. Farhadi, *Yolov3: An incremental improvement*, *arXiv preprint arXiv:1804.02767* (2018) .

[72] A. Mustapha, L. Mohamed and K. Ali, *An overview of gradient descent algorithm optimization in machine learning: Application in the ophthalmology field*, in *Smart Applications and Data Analysis*, M. Hamlich, L. Bellatreche, A. Mondal and C. Ordonez, eds., (Cham), pp. 349–359, Springer International Publishing, 2020.

[73] A. Bochkovskiy, C.-Y. Wang and H.-Y.M. Liao, *Yolov4: Optimal speed and accuracy of object detection*, *arXiv preprint arXiv:2004.10934* (2020) .

[74] H. Zhang, M. Cisse, Y.N. Dauphin and D. Lopez-Paz, *mixup: Beyond empirical risk minimization*, 1710.09412.