MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion

Shih-Yun Chu
National Taiwan University
r09922115@csie.ntu.edu.tw

Ming-Sui Lee
National Taiwan University
mslee@csie.ntu.edu.tw

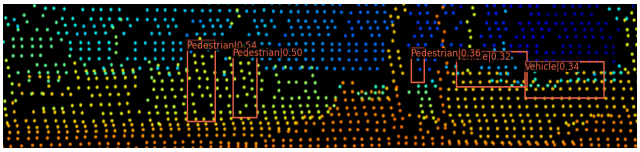
Abstract

Due to the trending need for autonomous driving, camera-based object detection has recently attracted lots of attention and successful development. However, there are times when unexpected and severe weather occurs in outdoor environments, making the detection tasks less effective and unexpected. In this case, additional sensors like lidar and radar are adopted to help the camera work in bad weather. However, existing multimodal detection methods do not consider the characteristics of different vehicle sensors to complement each other. Therefore, a novel end-to-end multimodal multistage object detection network called MT-DETR is proposed. Unlike the unimodal object detection networks, MT-DETR adds fusion modules and enhancement modules and adopts a hierarchical fusion mechanism. The Residual Fusion Module (RFM) and Confidence Fusion Module (CFM) are designed to fuse camera, lidar, radar, and time features. The Residual Enhancement Module (REM) reinforces each unimodal branch while a multistage loss is introduced to strengthen each branch's effectiveness. The synthesis algorithm for generating camera-lidar data pairs in foggy conditions further boosts the performance in unseen adverse weather. Extensive experiments on various weather conditions of the STF dataset demonstrate that MT-DETR outperforms state-of-the-art methods. The generality of MT-DETR has also been confirmed by replacing the feature extractor in the experiments. The code and pre-trained models are available on `https://github.com/Chushihyun/MT-DETR`.

1. Introduction

In the field of computer vision, object detection is a classic and valuable task [33, 6]. Its purpose is to find and classify bounding boxes of objects in an image, which can be applied to various scenarios. Nowadays, more accurate, faster, and lighter methods are still being proposed [7, 45]. With the development of automotive technology, object detection models have been widely used on the road [11, 5, 3, 39]. Safety is the most important thing when driving. For drivers, poor visibility at night or in heavy rain greatly increases the risk of driving. Under such conditions, the visibility of the camera is limited, so the camera-based detection model cannot assist drivers well. Fortunately, using multiple sensors can make up for the shortcoming of cameras. In this paper, not only an RGB camera but also lidar, radar, and time information are adopted as the model inputs. As shown in Fig.1, with the rich depth information of lidar and radar, more accurate and robust predictions are obtained so that the potential hazards on the road can be prevented and safety is guaranteed.

(a) Only camera is considered.



(b) The lidar data is too sparse.

(c) MT-DETR (proposed)

(d) Ground Truth

Figure 1. **Predictions of unimodal methods and MT-DETR.**

Considering the data type, information volume, and

characteristics of the sensors, the MT-DETR, an end-to-end multimodal multistage object detection model with confidence fusion, is proposed to tackle the object detection task in adverse weather. The followings are the contributions of this paper:

1. A novel end-to-end multimodal object detection model called MT-DETR is proposed. It includes fusion modules and enhancement modules and adopts a hierarchical fusion mechanism. Based on the importance of each modality, RFM, the confidence-aware CFM and REM are designed as fusion modules and enhancement module, respectively.

2. In order to ensure the effectiveness of each unimodal branch, MT-DETR adopts a multistage loss function, which increases stability and performance during training without causing extra time during inference.

3. The camera-lidar synthesis algorithm for foggy days (including day scenes and night scenes) is introduced by considering the glare effects and the variation of atmospheric light over time. The more realistic synthetic data make the model more adaptable to adverse weather during the training phase.

Extensive experiments on the STF dataset [3] demonstrate that MT-DETR outperforms the existing methods by a large margin both in clear and unseen adverse weather. If MT-DETR is additionally trained with the proposed synthetic foggy data, it achieves higher accuracy.

2. Related Work

2.1. Object Detection Architecture

Object detection is a popular and challenging task in computer vision. Most object detection frameworks can be roughly divided into one-stage and two-stage methods. One-stage detectors (e.g., SSD [24], RetinaNet [21], YOLO series [31, 32, 4, 42]) directly obtain the classification and bounding box of the object from the feature map quickly. In contrast, two-stage detectors (e.g., RCNN series [14, 13, 33, 16, 6, 8]) perform region proposal to propose candidate bounding boxes for better results with higher time cost.

Recently, [7] proposed a Transformer-based [40] end-to-end detection framework called DETR. It treats object detection as a set prediction problem, removing most of the previous hand-crafted design and becoming a simpler detection pipeline. The subsequent Deformable DETR [45] accelerated the convergence of DETR and achieved better performance. Based on this end-to-end framework, we propose MT-DETR using a multimodal backbone to process data from multiple sensors, and it is adaptive to almost all detection frameworks.

2.2. Multimodal Fusion

Due to the inherent insufficiency of a single sensor, multimodal fusion is a task that has received more and more attention recently. Many previous works [41, 29, 18] have studied how to combine images with text, sound or point clouds, etc. Compared with unimodal object detection (Fig.2(a)) [33, 31, 7, 45], Cross-modal data can be complementary for better performance. Multimodal fusion is a core but non-trivial problem in multimodal task. "Early Fusion" [41] simply concatenates multimodal inputs, and "Middle Fusion" [44, 3, 27, 23, 29] usually performs better by fusing cross-modal data at the feature extraction stage (Fig.2(b)). However, multimodal fusion is not a simple task because of the different data types and properties among the modalities. Existing multimodal object detection methods [44, 27, 23] successfully improve accuracy in clear weather but cannot perform well in bad weather. In this paper, MT-DETR is extended from "Middle Fusion" with additional specially designed modules and auxiliary loss functions (Fig.2(c)).

2.3. Datasets for Autonomous Vehicle

With the rapid development of autonomous driving technology, many camera-based driving datasets are presented [12, 11, 36]. Vehicles nowadays are usually equipped with multiple sensors (e.g., stereo cameras, lidar, radar) for more accurate detection, and many multimodal driving datasets are presented [5, 39, 3, 2, 10, 38]. For safe driving, autonomous vehicles should be able to adapt to any weather and time of day. Still, most driving datasets only focus on clear weather conditions [5, 39, 10].

The STF dataset [3] includes clear, light fog, dense fog, and snow weather, and each weather has day and night scenes. Due to the rich diversity of weather and time, it is adopted as the dataset for this research. Although STF covers a variety of weather, the amount of adverse weather data (light fog, dense fog, snow/rain) is insufficient for training. Therefore, we only use clear weather as training data and test on various weather conditions.

Considering commonality and effectiveness for general situations, data generated by the camera, lidar, radar, and time from STF are considered [3] as the input sensors. The followings are their characteristic:

- Camera has always been the primary sensor for object detection because it is the most abundant data source. However, the camera is limited when the visibility is not high due to insufficient light at night.

- Lidar is also an essential sensor for self-driving cars since it provides depth information and is not affected by brightness. Nevertheless, lidar visibility is reduced and noise occurs on rainy or foggy days.

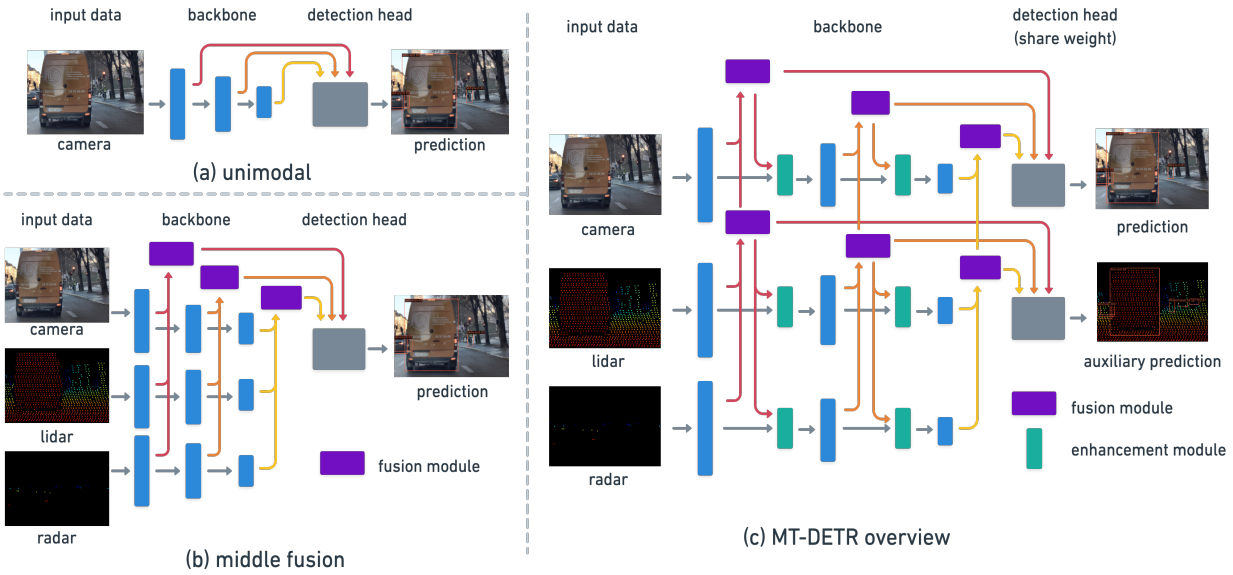


Figure 2. **Overview of different object detection frameworks.** (a) Unimodal methods input RGB camera image. After the feature extraction by backbone, multi-scale features are passed to detection head for prediction. (b) Middle fusion-based methods extract multimodal features by each branch, then fuse them together to predict objects. (c) MT-DETR uses fusion module and enhancement module for more accurate features, and adopts hierarchical fusion mechanism and auxiliary multistage loss to ensure more effective learning.

- Radar has good robustness and is not affected by bad weather, but the data points provided by radar in STF are very sparse. It is challenging to use radar effectively, especially with only dozens of depth points in a frame.
- Different from previous works [2, 5], we are the first to utilize time information as the model's input as we found that the reliability of each sensor may be influenced over time. Time data is a binary value according to the day/night annotation of the dataset so that the model can be aware of time information.

Due to the above reasons, it is believed that integrating information from cameras, lidar, radar, and time makes the model more adaptive to various weather conditions as those sensors contain complementary properties.

3. Methodology

3.1. MT-DETR

The proposed MT-DETR, a novel MulTimodal MulTistage network for end-to-end object detection, takes multiple sensors simultaneously with fused features. Same as the Deformable DETR [45], MT-DETR adopts Transformer [40] as the detection head. By fully utilizing each sensor, MT-DETR obtains robust detection results at night and in unseen weather. Experiments on STF dataset [3] demonstrate that MT-DETR outperforms unimodal and state-of-the-art methods.

3.1.1 Framework Overview

The input of MT-DETR's backbone is the image data obtained from different sensors, and the output is multi-scale features after multimodal fusion. Fig. 3 presents the architecture of MT-DETR, which consists of four components: feature extractor, fusion module, enhancement module, and detection head. The ConvNeXt [25] is adopted as the feature extractor in parallel unimodal branches. The fusion module fuses the features extracted by the unimodal branches (Section 3.2.1). The enhancement module combines the fused features with the unimodal branch to enhance the unimodal feature and then proceeds to the feature extraction of the next scale (Section 3.2.2). Finally, the fused features obtained from the fusion module at each scale are passed to the detection head for prediction.

3.1.2 Hierarchical Fusion Mechanism

Combining features from all branches simultaneously may lose the relationship and priority between sensors. Thus the hierarchical fusion mechanism is proposed. Because the data types of lidar and radar are similar, fusing them first can capture more thorough and accurate depth information. Therefore, when mixing modalities, we fuse lidar and radar into the depth feature, combining it with the camera and time branch. With the hierarchical fusion mechanism, the model can understand the depth of knowledge more clearly.

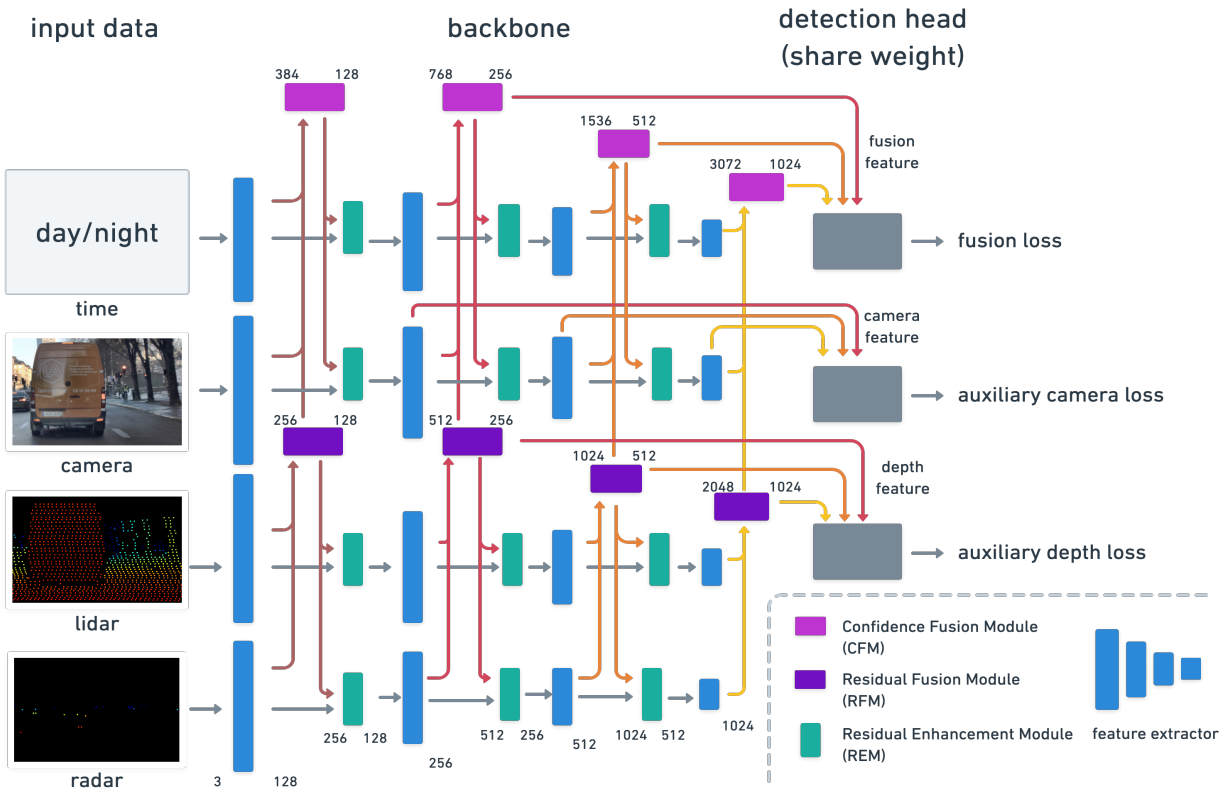


Figure 3. **The architecture of MT-DETR.**

3.2. Module Design

3.2.1 Fusion Modules

The Residual Fusion module (RFM) and Confidence Fusion Module (CFM) are proposed for the fusion purpose. As shown in Fig. 4(a)(b), RFM fuses the features of the lidar and radar branches to obtain the depth feature, and CFM is responsible for fusing the depth feature with camera and time branches to get the final fused features.

Taking the amount of information from lidar and radar into account, RFM concatenates the features of lidar and radar. Then its dimension is reduced by the convolution block, which adds to the lidar features' residual connection. Considering the characteristics of each modal, CFM concatenates the features of the camera, depth, and time to obtain the confidence map after dimensionality reduction by the convolution block. After that, the confidence map is element-wisely multiplied with the depth feature and added to the camera feature, becoming the fusion feature.

The depth feature $\mathbf{F}_i^{\text{depth}}$ and fusion feature $\mathbf{F}_i^{\text{fusion}}$ of i-th stage can be computed by:

$$\begin{aligned}\mathbf{F}_i^{\text{depth}} &= \text{RFM}(\mathbf{F}_i^{\text{lidar}}, \mathbf{F}_i^{\text{radar}}) \\ &= \mathbf{F}_i^{\text{lidar}} + \text{Conv}_{1\times1}(\mathbf{F}_i^{\text{lidar}} \oplus \mathbf{F}_i^{\text{radar}}) \,,\end{aligned} \tag{1}$$

$$\begin{aligned}\mathbf{F}_i^{\text{fusion}} &= \text{CFM}(\mathbf{F}_i^{\text{camera}}, \mathbf{F}_i^{\text{depth}}, \mathbf{F}_i^{\text{time}}) \\ &= \mathbf{F}_i^{\text{camera}} + (\mathbf{F}_i^{\text{depth}} * \sigma(\text{Conv}_{1\times1}(\mathbf{F}_i^{\text{camera}} \oplus \mathbf{F}_i^{\text{depth}} \oplus \mathbf{F}_i^{\text{time}}))) \,,\end{aligned} \tag{2}$$

where $\mathbf{F}_i^{\text{camera}}, \mathbf{F}_i^{\text{lidar}}, \mathbf{F}_i^{\text{radar}}, \mathbf{F}_i^{\text{time}}$ denote the feature outputted from i-th stage ($i = 1, 2, 3, 4$) of each unimodal feature extractor, $\oplus$ denotes an operation of feature concatenation, $*$ and $+$ denote element-wise multiplication and addition respectively, $\sigma(\cdot)$ and $\text{Conv}_{1\times1}(\cdot)$ indicate sigmoid function and 1×1 convolution block.

3.2.2 Enhancement Module

The Residual Enhancement Module (REM) is proposed to reinforce the unimodal branch. As shown in Fig. 4(c), REM is similar in structure to RFM but has a deeper convolution layer. In addition, REM pays more attention to the features of the unimodal branch. That is to say, it combines the output of the convolution block with the unimodal feature to be the enhanced feature. It should be noted that the camera and time branches are enhanced using the fusion feature, while the lidar and radar branches are boosted with the

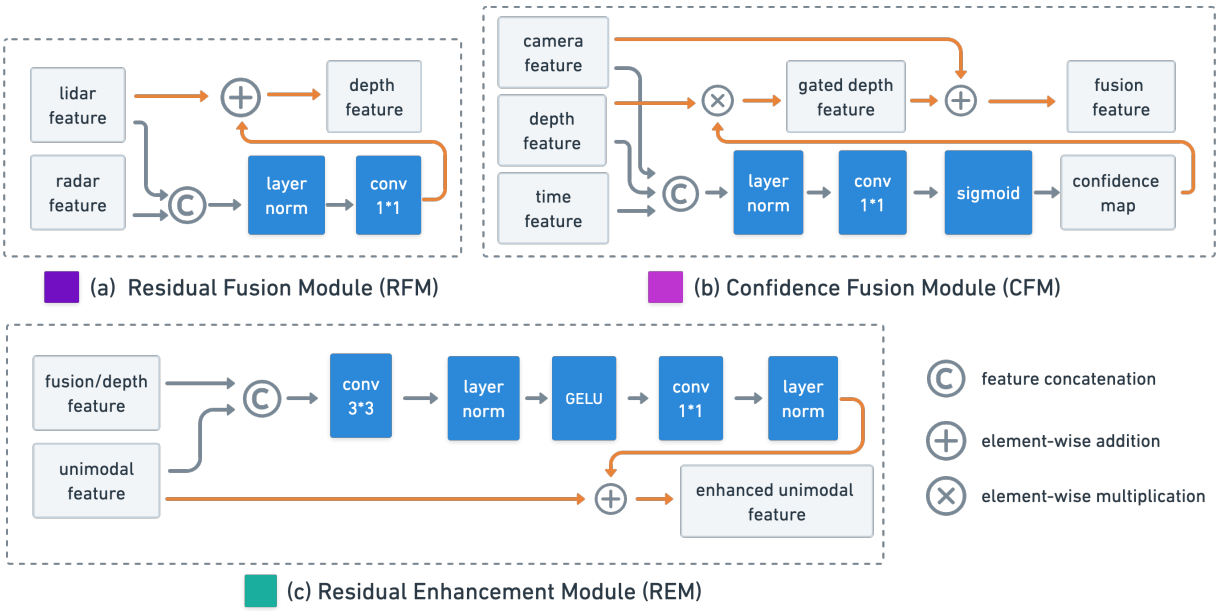


Figure 4. **The architecture of fusion modules and enhancement module.** (a) Residual Fusion module (RFM) fuses the features of lidar and radar into depth feature. (b) Confidence Fusion Module (CFM) fuses the features of camera, depth, and time to become the final fusion feature. (c) Residual Enhancement Module (REM) fuses unimodal feature with fusion feature into refined unimodal feature.

depth feature. The enhanced feature $\widetilde{\mathbf{F}}_i^m$ of each unimodal branch can be obtained by:

$$\begin{aligned}\widetilde{\mathbf{F}}_i^m &= \mathrm{REM}_m(\mathbf{F}_i^m, \mathbf{F}_i^{\text{fusion}}),\ \text{for } m \in \{\text{camera, time}\} \\ &= \mathbf{F}_i^m + \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(\mathbf{F}_i^m \oplus \mathbf{F}_i^{\text{fusion}}))\ ,\end{aligned} \tag{3}$$

$$\begin{aligned}\widetilde{\mathbf{F}}_i^m &= \mathrm{REM}_m(\mathbf{F}_i^m, \mathbf{F}_i^{\text{depth}}),\ \text{for } m \in \{\text{lidar, radar}\} \\ &= \mathbf{F}_i^m + \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(\mathbf{F}_i^m \oplus \mathbf{F}_i^{\text{depth}}))\ ,\end{aligned} \tag{4}$$

then the feature at the next scale $\mathbf{F}_{i+1}^m$ can be extracted as:

$$\mathbf{F}_{i+1}^m = \mathrm{FE}_{i+1}^m(\widetilde{\mathbf{F}}_i^m),\ \text{for } i \in \{1, 2, 3\}\ , \tag{5}$$

where $\mathrm{REM}_m(\cdot,\cdot)$ and $\mathrm{FE}_i^m(\cdot)$ indicate the REM module and feature extractor of each unimodal branch at i-th stage.

For $m \in \{\text{fusion, camera, depth}\}$, we finally collect $\mathbf{F}_m = \{\mathbf{F}_i^m \mid i = 2, 3, 4\}$ and feed them into the following detection head for further prediction.

3.3. Multistage Loss Function

While fusing with extremely unbalanced sensor information, it is possible for the fusion module only to trust the features from the camera branch. Therefore, the rest of the unimodal branches may not be well learned. Inspired by [20, 6], the final fusion feature and middle-stage features are fed into the head to get the detection result for calculating the auxiliary loss function. This multistage loss function ensures that each branch can extract useful information. For unimodal object detection, the loss function is the same as Deformable DETR [45]: Focal loss [21] for classification, l_1 loss and Generalized IoU loss [34] for bounding box regression. These loss functions are added to follow the weights of the Deformable DETR [45].

Afterwards, we use the prediction obtained from $\mathbf{F}_{\text{fusion}}$ to calculate the fusion loss $\mathcal{L}_{\text{fusion}}$; the prediction obtained from $\mathbf{F}_{\text{camera}}$ to calculate the camera loss $\mathcal{L}_{\text{camera}}$; and the prediction obtained from $\mathbf{F}_{\text{depth}}$ to calculate the depth loss $\mathcal{L}_{\text{depth}}$. $\mathcal{L}_{\text{fusion}}$ is the primary loss while $\mathcal{L}_{\text{camera}}$ and $\mathcal{L}_{\text{depth}}$ are auxiliary ones. The total loss is then defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{fusion}}\mathcal{L}_{\text{fusion}} + \lambda_{\text{camera}}\mathcal{L}_{\text{camera}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}\ , \tag{6}$$

where λ_{fusion}, λ_{camera}, λ_{depth} are corresponding weights to balance the assistant supervision.

3.4. Foggy Data Synthesis

In the STF dataset, only clear weather data are considered for training, so we propose a synthesis method to generate foggy data from those clear data. A pair of data contains information from the camera, lidar, radar, and time. Notice that the data from the camera and lidar are affected by fog, so we attempt to generate the same density of fog on both the camera and lidar data.

3.4.1 Camera Data Synthesis

Based on [19, 35], the following composition formulas are utilized to impose fog on clear images. Given clear image I, depth map D, atmosphere light A, the transmission map T, the foggy image $I^{'}$ can be generated as:

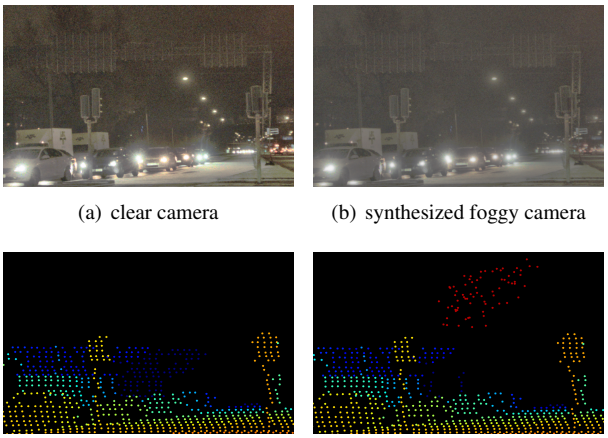
$$\begin{aligned} T &= e^{-\beta \times D}\,, \\ I^{'} &= T * I + (1 - T) * A\,, \end{aligned} \tag{7}$$

where β is a weight representing the density of fog, which is set to 1.0, $*$ and + denote pixel-wise multiplication and addition, respectively.

As no depth map is available in STF, the pretrained depth estimation model DPT [30] is adopted to predict the depth. In [19], the atmosphere light A is a randomly sampled number from $[0.3, 0.7]$ for the whole frame, yet we introduce two modifications. (1) The sampling interval is adjusted over time to consider the difference between day and night. (2) Real foggy images have apparent glare effects, so the atmosphere light A varies by referring to the local brightness of the camera image I. In other words, the atmosphere light A is rather than a single value for the whole frame. The effect of improved fog synthesis is shown in Fig. 5(b). More details and visualizations of the synthesis algorithm are provided in the supplementary material.

3.4.2 Lidar Data Synthesis

Due to the limited penetration of the invisible light emitted by the lidar sensor, the sensing ability of the lidar is affected in adverse weather. For example, fog affects the lidar signal in two aspects. First, visibility becomes poor, and lidar points in far distance will disappear. Second, noise is introduced into the air. Based on the point cloud imaging principle of lidar, [15] physically accurately simulated the effect of fog on lidar and proposed a fog simulation algorithm on clear lidar data. We utilize the method of [15] to add the fog effect on the lidar point cloud and the resultant effect is shown in Figure 5(d).



(a) clear camera (b) synthesized foggy camera

(c) clear lidar (d) synthesized foggy lidar

Figure 5. **Results of fog synthesis on camera and lidar.** (b) is synthesized from (a) using equation (7), and (d) is synthesized from (c) using the method proposed by [15].

Table 1. **The amount of each weather condition in STF.**

Condition	Training	Validation	Testing			
	clear	clear	clear	light fog	dense fog	snow
day	2183	399	1005	633	572	2293
night	1343	409	877	419	315	2440
total	3526	808	1882	1052	887	4733

4. Experiment

4.1. Dataset, Metrics and Implementation Details

The experiments are conducted on the STF dataset [3]. It provides data with 2D bounding boxes for object detection of vehicles and pedestrians in various weather conditions, including clear, light fog, dense fog, and snow. All training, validation, and testing data include day and night scenarios, along with their weather conditions and quantities are shown in Table 1.

The average precision (AP) is used to evaluate the object detection tasks. As for the COCO [22] benchmark, AP_{75} is adopted as the scoring metric. AP_{75} represents the area under the precision-recall curve, so it falls between 0 and 100 (%). The higher the score, the more accurate the model.

For a fair comparison, all models adopt ConvNeXt [25] as the feature extractor and the detection head of Deformable DETR [45] and follow the same training settings and parameters. All models (unless otherwise noted) are trained on clear weather data only. We implement MT-DETR and reimplement the previous methods in mmdetection toolbox [9] with Pytorch [28]. All experiments are conducted on a single Nvidia A6000 GPU. AdamW optimizer [26] is utilized to train MT-DETR for 36 epochs with batch size 1. The learning rate starts at 0.0001 with layer-wise learning rate decay [1] and the weight decay is set to be 0.05. $\lambda_{\text{fusion}} = 1.0$, $\lambda_{\text{camera}} = 1.0$, and $\lambda_{\text{depth}} = 0.5$ are chosen to balance the multistage loss.

4.2. Multimodal Effectiveness

The performance of each combination of sensors in MT-DETR is shown in Table 2. The results verify that the fusion of three sensors outperforms a single or a mixture of two, ensuring that MT-DETR integrates the advantages of each sensor successfully. For unimodal comparisons, the camera performs better than lidar and far exceeds radar due to the amount of information. Radar's points provided by STF are too sparse to make successful predictions.

The comparison of camera, camera-lidar, and camera-radar demonstrates that lidar and radar can assist the camera. Since lidar has more information, camera-lidar performs better than camera-radar in clear weather. However, because radar signal has better penetration, camera-radar can catch up with camera-lidar performance in adverse weather conditions. The camera-lidar-radar one surpasses any of the above models, which confirms that the three sensors' information can complement each other, and MT-DETR effectively utilizes their properties to improve accuracy.

Table 2. **The performance of different inputs on all STF test splits.** The best and second best results are highlighted in **bold** and underlined, respectively.

Training Data			Testing Data							
			clear		light fog		dense fog		snow	
camera	lidar	radar	day	night	day	night	day	night	day	night
✓			62.1	58.9	63.4	59.9	69.6	67.9	63.0	61.1
	✓		27.9	32.6	19.3	33.6	19.5	16.1	28.2	30.2
		✓	0.8	0.6	0.9	0.4	0.8	0.5	0.8	0.4
✓	✓		64.0	60.7	64.8	62.9	69.9	69.4	64.5	63.8
✓		✓	63.7	60.1	**66.4**	61.8	70.3	69.4	64.5	63.1
✓	✓	✓	**65.0**	**61.8**	66.2	**63.3**	**71.5**	**69.6**	**65.4**	**64.2**

4.3. Comparison

In this section, MT-DETR is compared with baselines and state-of-the-art methods. Since some existing methods take the camera-radar signals as input, while some serve the camera-lidar-radar signal to the model, the performance comparison is provided with these two settings. Moreover, experiments are conducted to verify whether the proposed time branch and the proposed synthetic training data are beneficial for the model's performance. More visual predictions of MT-DETR are shown in the supplementary material.

4.3.1 Camera-radar

In this section, only the camera and radar signals are fed into the MT-DETR. In Table 3, the MT-DETR is compared with state-of-the-art object detection methods [27, 44]. Two baselines are considered here: the "Early Fusion" method feeds the concatenation of each sensor into a unimodal model at the beginning; the "Middle Fusion" method (like Fig. 2(b)) fuses the features extracted by each branch and sends them to the detection head. Camera-radar fusion is challenging because radar data is too sparse in the STF dataset. As it can be seen from Table 3, the proposed approach achieves higher performance than other methods in all weather conditions. While other methods are easily overlooked or misled by radar, MT-DETR can leverage radar well to surpass them.

Table 3. **Comparison of the baselines and state-of-the-art methods with camera-radar signals on all STF test splits.** The best and second best results are highlighted in **bold** and underlined, respectively.

Method	Testing Data							
	clear		light fog		dense fog		snow	
	day	night	day	night	day	night	day	night
Early Fusion	61.5	56.8	58.1	60.7	60.7	60.7	60.6	61.0
Middle Fusion	61.6	58.5	64.3	60.3	69.2	67.3	63.0	61.1
CRFNet [27]	62.3	57.7	62.5	60.3	68.3	67.7	62.3	60.9
BiRANet [44]	61.8	57.7	60.4	60.8	69.0	68.0	62.0	61.1
MT-DETR (Ours)	**63.7**	**60.1**	**66.4**	**61.8**	**70.3**	**69.4**	**64.5**	**63.1**

4.3.2 Camera-lidar-radar

Here the camera-lidar-radar signals serve as input to the MT-DETR in the experiments. The comparisons with baselines and state-of-the-art multimodal methods [23, 3] are provided in Table 4. The baselines and state-of-the-art methods are fairly comparable under different weather conditions, while it is clear that the proposed MT-DETR outperforms other methods by noticeable margins. These results further prove that the design of MT-DETR, such as CFM and multistage loss, can improve performance in clear weather and enhance robustness in adverse environments.

Table 4. **Comparison of baselines and state-of-the-art methods with camera-lidar-radar signals on all STF test splits.** The best and second best results are highlighted in **bold** and underlined, respectively.

Method	Testing Data							
	clear		light fog		dense fog		snow	
	day	night	day	night	day	night	day	night
Early Fusion	61.9	59.1	60.7	61.8	57.8	60.2	62.0	61.8
Middle Fusion	63.4	59.6	62.1	62.0	69.1	67.7	64.3	62.4
IADM [23]	60.4	57.4	60.3	59.6	67.5	67.0	60.4	59.2
DEF [3]	62.9	59.8	65.6	61.9	69.4	69.2	64.1	62.6
MT-DETR (Ours)	**65.0**	**61.8**	**66.2**	**63.3**	**71.5**	**69.6**	**65.4**	**64.2**

4.3.3 Overall System

Here we propose introducing the time branch and using the proposed synthetic foggy data for training. As illustrated in Table 5, the MT-DETR with camera-lidar-radar-time signals (the full model) performs well in various weather conditions. In addition, the full model with the proposed synthetic training data boosts the results for almost all the cases except the night scene with dense fog, which is a challenging case that requires further discussion. Here we also consider the glare effect in the generating process of the synthetic data. The improvement of considering the glare effect is also reflected in the results. The glare effect generation is detailed in the supplementary material.

Table 5. **The full MT-DETR with synthetic training data.** The best and second best results are highlighted in **bold** and underlined, respectively. * denotes the synthesis algorithm without glare effect.

Training Data					Testing Data							
camera	lidar	radar	time	synthetic data	clear day	clear night	light fog day	light fog night	dense fog day	dense fog night	snow day	snow night
✓	✓	✓			65.0	61.8	66.2	63.3	71.5	69.6	65.4	64.2
✓	✓	✓	✓		64.7	62.2	67.0	63.7	71.3	**70.7**	65.9	64.8
✓	✓	✓	✓	✓*	65.7	**63.2**	67.2	65.3	70.3	68.6	66.9	**65.6**
✓	✓	✓	✓	✓	**66.2**	63.1	**68.0**	**65.8**	**71.7**	70.1	**67.2**	**65.6**

4.4. Ablation Study

4.4.1 Ablation on MT-DETR Input Sensor

To verify that every sensor provides helpful information and improves the model's performance, we conduct an ablation study on the input sensors. Table 6 shows that MT-DETR performs best when using all sensors simultaneously (camera, lidar, radar, time). Considering time information, although the full MT-DETR model's performance is not as expected in cases with clear days and days with dense fog, it provides better performance in other conditions. Generally speaking, each branch provides an improvement differently in different cases.

Table 6. **Ablation study of MT-DETR input sensors.** The best result is highlighted in **bold**.

Input Sensor				Testing Data							
camera	lidar	radar	time	clear day	clear night	light fog day	light fog night	dense fog day	dense fog night	snow day	snow night
	✓	✓	✓	28.8	34.1	19.3	35.0	19.0	18.2	29.6	31.2
✓		✓	✓	63.6	59.6	66.0	62.2	70.6	69.4	64.1	62.5
✓	✓		✓	64.3	61.2	64.7	63.2	70.8	69.8	65.6	64.4
✓	✓	✓		**65.0**	61.8	66.2	63.3	**71.5**	69.6	65.4	64.2
✓	✓	✓	✓	64.7	**62.2**	**67.0**	**63.7**	71.3	**70.7**	**65.9**	**64.8**

4.4.2 Ablation of MT-DETR Architecture

An ablation study to investigate the effectiveness of the proposed components of MT-DETR is provided in this section. As shown in Table 7, the "Early Fusion" concatenates the data of each sensor at the beginning, so there is no multimodal design which means no fusion is considered. "Middle Fusion" fuses the features by the simplest approach. "w/o REM" removes the enhancement module from the model, which means the fusion feature is not merged with any branch. "w/o CFM" replaces CFM with simpler RFM. "w/o Hierarchical" fuses all sensors at once instead of fusing lidar and radar first. "w/o Multistage" only passes the fusion feature to the following detection head, so there are no auxiliary loss functions $\mathcal{L}_{\text{camera}}$ and $\mathcal{L}_{\text{depth}}$. These results confirm the effectiveness of every proposed design. An ablation study on synthetic fog density is reported in the supplementary material.

Table 7. **Ablation study of MT-DETR components.** The best result is highlighted in **bold**.

Method	Testing Data							
	clear day	clear night	light fog day	light fog night	dense fog day	dense fog night	snow day	snow night
Early Fusion	59.5	57.3	54.4	60.0	57.5	54.7	59.2	59.7
Middle Fusion	63.6	59.9	64.7	62.8	68.5	69.0	64.0	62.9
w/o REM	63.5	61.0	65.4	62.8	71.1	68.7	65.3	64.0
w/o CFM	**64.7**	62.0	64.9	63.4	70.1	68.9	65.5	64.5
w/o Hierarchical	63.1	59.9	63.9	62.0	70.7	68.2	64.1	62.5
w/o Multistage	62.9	60.3	64.7	62.7	69.2	68.4	64.2	62.6
Full MT-DETR (ours)	**64.7**	**62.2**	**67.0**	**63.7**	**71.3**	**70.7**	**65.9**	**64.8**

4.5. Generalization to Other Feature Extractors

The feature extractor of MT-DETR can be flexibly replaced to meet different needs. Table 8 shows the results with different feature extractors. The ConvNeXt-b [25] is replaced by ResNet-50 [17], ResNeXt-101 [43] and MobileNetV2 [37], respectively. The middle fusion is the baseline, and it is compared to the proposed MT-DETR. As can be seen from the table, the MT-DETR performs better in all settings, which confirms its generality.

Table 8. **The MT-DETR and middle fusion with different feature extractors.** The better results are highlighted in **bold**.

Model Architecture		Testing Data							
Feature Extractor	Fusion Method	clear day	clear night	light fog day	light fog night	dense fog day	dense fog night	snow day	snow night
ConvNeXt-b [25]	Middle Fusion	63.6	59.9	64.7	62.8	68.5	69.0	64.0	62.9
	MT-DETR	**64.7**	**62.2**	**67.0**	**63.7**	**71.3**	**70.7**	**65.9**	**64.8**
ResNet-50 [17]	Middle Fusion	59.8	57.7	55.3	59.6	54.4	53.9	58.4	56.0
	MT-DETR	**61.4**	**59.3**	**57.9**	**61.5**	**57.2**	**57.1**	**61.3**	**61.4**
ResNeXt-101 [43]	Middle Fusion	59.5	57.1	54.7	58.9	55.9	**53.4**	59.3	59.6
	MT-DETR	**61.1**	**58.5**	**56.4**	**61.3**	**56.0**	51.2	**61.2**	**61.2**
MobileNetV2 [37]	Middle Fusion	27.3	52.8	26.0	55.1	28.7	39.0	28.2	53.9
	MT-DETR	**57.4**	**54.1**	**51.1**	**56.0**	**52.0**	**41.2**	**56.0**	**55.0**

5. Conclusion

This paper proposes a novel MT-DETR network for multimodal object detection, which adopts RFM, CFM, REM, and hierarchical fusion mechanism to perform cross-modal fusion and exchange. Additionally, MT-DETR employs a multistage loss to address the imbalance among vehicle sensors and learn to extract compelling features. MT-DETR achieves state-of-the-art performance using the camera, lidar and radar, and even better with additional time information and the proposed synthetic foggy training data. The experimental results demonstrate that the MT-DETR is robust and performs well in various weather conditions. The good generalization and scalability confirm future applicability to different multimodal tasks.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[2] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6433–6438. IEEE, 2020.

[3] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020.

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[8] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.

[9] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[10] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[15] Martin Hahner, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15283–15292, 2021.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[19] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1633–1642, 2019.

[20] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cbnetv2: A composite backbone network architecture for object detection. *arXiv preprint arXiv:2107.00420*, 2021.

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, and Liang Lin. Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4823–4833, 2021.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[25] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[27] Felix Nobis, Maximilian Geisslinger, Markus Weber, Johannes Betz, and Markus Lienkamp. A deep learning-based radar and camera sensor fusion architecture for object detection. In *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pages 1–7. IEEE, 2019.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[29] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7077–7087, 2021.

[30] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[32] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[34] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 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*, pages 658–666, 2019.

[35] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the european conference on computer vision (ECCV)*, pages 687–704, 2018.

[36] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021.

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[38] Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. Radiate: A radar dataset for automotive perception in bad weather. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2021.

[39] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Jörg Wagner, Volker Fischer, Michael Herman, Sven Behnke, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*, volume 587, pages 509–514, 2016.

[42] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 13029–13038, 2021.

[43] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[44] Ritu Yadav, Axel Vierling, and Karsten Berns. Radar+ rgb attentive fusion for robust object detection in autonomous vehicles. *arXiv preprint arXiv:2008.13642*, 2020.

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.