Enhancing Object Detection in Adverse Weather Conditions through Entropy and Guided Multimodal Fusion

Zhenrong Zhang[1 [0009−0005−8570−4334]], Haoyan Gong[1 [0000−0002−6572−6174]], Yuzheng Feng[1 [0009−0001−9884−5807]], and Zixuan Chu[1 [0009−0007−3670−7461]] and Hongbin Liu[1 ⋆[0009−0008−3041−370X]]

Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China Zhenrong.Zhang21, Haoyan.Gong21, Yuzheng.Feng21, Zixuan.Chu23@student.xjtlu.edu.cn hongbin.liu@xjtlu.edu.cn

Abstract. Integrating diverse representations from complementary sensing modalities is essential for robust scene interpretation in autonomous driving. Deep learning architectures that fuse vision and range data have advanced 2D and 3D object detection in recent years. However, these modalities often suffer degradation in adverse weather or lighting conditions, leading to decreased performance. While domain adaptation methods have been developed to bridge the gap between source and target domains, they typically fall short because of the inherent discrepancy between the source and target domains. This discrepancy can manifest in different distributions of data and different feature spaces. This paper introduces a comprehensive domain-adaptive object detection framework. Developed through deep transfer learning, the framework is designed to robustly generalize from labelled clear-weather data to unlabeled adverse weather conditions, enhancing the performance of deep learning-based object detection models. The innovative Patch Entropy Fusion Module (PEFM) is central to our approach, which dynamically integrates sensor data, emphasizing critical information and minimizing background distractions. This is further complemented by a novel Weighted Decision Module (WDM) that adjusts the contributions of different sensors based on their efficacy under specific environmental conditions, thereby optimizing detection accuracy. Additionally, we integrate a domain align loss during the transfer learning process to ensure effective domain adaptation by regularizing the feature map discrepancies between clear and adverse weather datasets. We evaluate our model on diverse datasets, including ExDark (unimodal), Cityscapes (unimodal), and Dense (multimodal), where it ranks 1^{st} in all datasets at the point in time of our evaluation.

Keywords: Multimodal fusion · Domain adaptation · Entropy fusion.

⋆ corresponding author

1 Introduction

Recently, advancements in deep learning have significantly contributed to autonomous driving, particularly in environmental perception tasks such as object detection and segmentation. These foundational tasks are crucial, as they represent the initial stages of the autonomous driving pipeline and must satisfy three critical conditions: accuracy, robustness, and real-time performance. While most existing models for environmental perception in autonomous driving are trained on high-quality data, their performance often deteriorates under adverse weather conditions such as low light or fog. This is primarily due to their reliance on clear weather datasets, which lack the variability necessary for robust performance in less ideal conditions [3].

(a) Camera Data (b) Gated Camera (c) Thermal Data (d) LIDAR Data

Fig. 1: This figure illustrates the data collected from different sensors under adverse weather conditions, each represented in separate subfigures.

As illustrated in Fig. 1, different sensors capture various information. This figure shows the applicability of different sensor types. For example, subfigure (a) displays an image captured by the RGB camera, which fails to capture the colour, texture, and detailed shapes of objects under adverse weather conditions. Subfigure (b) shows an image from the gated camera, which is designed to enhance visibility through adverse weather by filtering specific light frequencies. Subfigure (c) presents a thermal map, which highlights heat signatures and is crucial for identifying living objects in low visibility conditions. Subfigure (d) depicts the point cloud from LIDAR, offering precise three-dimensional information about the environment by measuring distances using laser light.

In order to improve the performance of object detections under adverse weather conditions, some efforts have been made to fine-tune pre-trained models on domain-specific datasets tailored to adverse conditions. However, this approach can be costly and may result in degraded performance when the models are applied back to their original domains. As an alternative, advanced image enhancement techniques have been developed to improve object detection by refining the multi-scale information within images and other sensor data [2, 36]. These techniques are particularly effective in enhancing the quality of visual data, which enhances the reliability of detection algorithms. However, despite their benefits, such enhancements focus on RGB camera data. This focus results in a significant underutilization of the broader potential offered by integrating multiple sensor modalities. Sensor fusion, which combines data from various sensors like RGB cameras, LIDAR, and radar, could harness a more synergistic

potential, offering a comprehensive view of the environment that surpasses the capabilities of any single sensor type.

Notably, the integration of data from multiple sensors, such as RGB cameras and active range sensors like LIDAR and radar, can enhance accuracy and robustness. These sensors provide complementary information about the vehicle's surroundings [11]. This method excels by integrating inputs from various sensors simultaneously, resulting in generally more models like [20,28] highlighting these advancements. However, those works do not consider that different sensors have different contributions in the feature fusion, and in the real world, situations exist that are only unimodal, which have a lot of noise in the background.

To find a method that can guide the multi-sensor fusions under adverse weather conditions and, if there are unimodal situations, how to enhance the feature map quality and eliminate the background noise is our motivation. Inspired by previous works like [3,11,29] in domain adaptation and sensor fusion, we have developed a novel model that innovatively combines a detection-driven enhancement network with an entropy-based fusion strategy. This model is designed to achieve a more profound integration of sensor data, which is particularly beneficial under various adverse weather conditions.

At the core of our approach is the entropy-steered deep fusion mechanism, which dynamically controls the fusion process by measuring entropy. This allows the model to adaptively emphasize relevant features and minimize the impact of background noise, which is prevalent in adverse weather conditions. These conditions typically degrade detection performance by introducing significant noise. By utilizing local measurement entropy, which assigns higher values to image regions with greater uncertainty, our mechanism ensures that only the most pertinent and reliable data are considered.

Furthermore, we have also incorporated a weighted decision module into our system, which assesses the effectiveness of each sensor's data based on current weather conditions and adjusts their contributions accordingly. To further develop robust, we design a new loss function based on Domain Adaptation (DA) methods [4, 22], which is fundamentally a learning paradigm designed to enhance generalization to out-of-domain data by transferring knowledge from a well-labelled source domain to an unlabeled target domain like [7,28,38]. Overall, the contributions of this paper are summarized as follows:

- We introduce a novel deep transfer learning-based domain adaptive object detection framework designed for autonomous driving under varying weather conditions.
- We propose the ***P***atch ***E***ntropy ***F***usion ***M***odule named ***PEFM***, which facilitates adaptive feature integration leveraging entropy that assigns higher values to regions with greater uncertainty under adverse weather and a ***W***eighted ***D***ecision ***M***odule named ***WDM***, that dynamically adjusts the contributions of various sensors based on their effectiveness under specific weather conditions. We also introduce a new domain align loss that applies during the transfer learning phase.

– Compared with different types of state-of-the-art (SOTA) methods, the proposed method can provide the most faithful detection results under adverse weather conditions in both ExDark, Cityscapes and Dense datasets.

2 Related Work

In this section, we provide a comprehensive review of the existing literature on object detection, focusing on different weather conditions—both clear and adverse. Additionally, we will explore the field of domain adaptation in object detection, which seeks to bridge the performance gap that often exists between these varying conditions.

2.1 Object detection in autonomous driving under clear weather

Recent advancements in deep learning have significantly propelled the field of autonomous driving, with notable progress in key areas such as object detection and segmentation [16,27]. Within this domain, object detection has emerged as a particularly active area of research [12]. Over the past decade, significant efforts have been directed towards enhancing both unimodal like [15] and multimodal 2D object detection systems under clear weather conditions like [10,26,31]. Key developments such as SSD [24], YOLOv3 [30], MV3D [6], and UAU [17] have demonstrated exceptional performance, setting new benchmarks for speed, accuracy, and robustness in autonomous driving scenarios. These models, however, are predominantly optimized for ideal weather conditions and often do not account for the complexities and challenges presented by adverse weather. The presence of fog, heavy rain, or snow can severely distort visual information, presenting substantial hurdles that are not typically addressed by standard object detection algorithms. This gap highlights a critical area of need within the field, as real-world driving conditions frequently involve navigating through less-than-perfect weather scenarios [3]. In response to this challenge, our work shifts focus towards developing robust object detection methods specifically tailored for adverse weather conditions. By addressing the shortcomings of existing models in handling environmental distortions, our research aims to enhance the reliability and safety of autonomous vehicles across a broader range of operating conditions, ensuring effective performance even in the face of severe weather disturbances.

2.2 Object detection in autonomous driving under adverse weather

Recent advancements in object detection under adverse weather conditions have largely centered on camera-only strategies. For instance, a data enhancement approach is presented that capitalizes on extracting multi-frequency information from a single image in [29]. Although this method shows effectiveness, its reliance on camera input alone makes it vulnerable to the challenges posed by poor visibility. In response to the limitations of camera-only systems, multimodal approaches have been developed to harness the complementary strengths

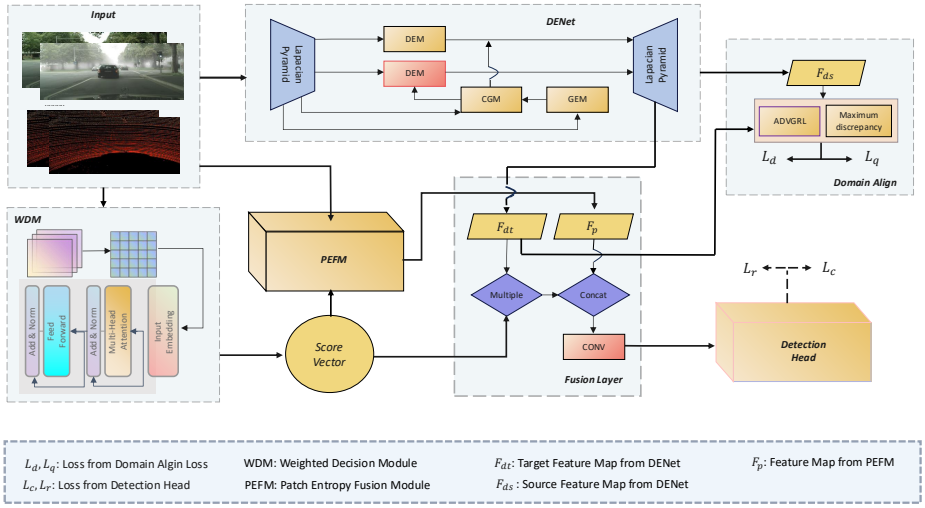

Fig. 2: This figure illustrates the architecture of our integrated framework.

of diverse sensors, including LIDAR, Radar, and RGB cameras. These strategies like [35, 39] aim to enhance overall detection performance, particularly in difficult weather conditions. In response to the challenges posed by foggy conditions, an innovative approach is developed in [35] by constructing a synthetic fog dataset. This dataset leverages the depth information from images within a virtual scene to simulate realistic fog scenarios. Building on this foundation, they have introduced an enhancement to the YOLOv5 architecture, specifically optimized for object detection in fog. The work in [11] integrates LIDAR and camera data through a fusion strategy that significantly boosts object detection capabilities in adverse environments. However, this approach primarily utilizes entropy weighting to determine the contribution of each modality based on its information content, a technique that, while innovative, can be further optimized. Building on these established methodologies, our work introduces two innovative modules that enhance the utilization of sensor data, drawing upon the foundational techniques of the work in [29]. Our model not only excels in unimodal scenarios but also significantly advances multimodal fusion capabilities. By doing so, it ensures robust performance across a wide spectrum of environmental challenges, thereby expanding the operational envelope for autonomous driving systems in adverse weather conditions.

3 Method

In this section, we will first introduce the definition and overall network architecture, then describe the WDM and PEFM modules, and finally, reveal the details of our loss function.

3.1 Definition

In our formulation, we define the set $\mathcal{N} = [\mathcal{N}_1, \ldots, \mathcal{N}_j, \ldots, \mathcal{N}_J]$ to represent a collection of labelled source data pairs, where each $\mathcal{N}_j = [\mathcal{I}_j, \mathcal{P}_j, \mathcal{Y}_j]$. $\mathcal{N}_j$ corresponds to a tuple containing the input image $\mathcal{I}_j$, the associated sensor data $\mathcal{P}_j$, and the ground truth labels $\mathcal{Y}_j$. $\mathcal{Y}_j$ comprises the instance labels for each detected object within the scene, characterized by bounding boxes and corresponding class labels, providing comprehensive semantic information necessary for subsequent perception tasks. If the input only has RGB camera image data, the $\mathcal{P}_j$ is set as an empty set.

3.2 Network Architecture

In our approach, we have introduced several key modules in our framework. Our framework architecture is shown in Fig. 2.

First, the data is input into WDM, which dynamically assigns weights to each sensor input. This module recognizes that different sensors contribute variably valuable information and noise depending on the prevailing weather conditions, allowing for adaptive data processing. WDM processes the inputs and determines the optimal weight for each sensor, ensuring that the fusion process is tailored to the specific environmental context and then PEFM is introduced, which selectively amalgamates the most pertinent data to enhance the reliability and redundancy of the information, facilitating more effective fusion. PEFM is ingeniously conceived with an entropy-steered deep fusion mechanism. It autonomously selects and amalgamates the most pertinent data to extract redundant and reliable information, particularly under challenging weather conditions. Through entropy folding and sigmoid transformation, a multiplication matrix is generated within the range $[0, 1]$, enabling dynamic scaling of the concatenated features from each sensor. The resultant entropy map, post-convolution and sigmoid activation to get the most informative features while mitigating the influence of non-essential background elements.

Moreover, the Detection-Driven Enhancement Network (DENet) framework in [29] is utilized, which bifurcates the input data into a low-frequency (LF) component through a Global Enhancement Module (GEM) and multiple high-frequency (HF) components via a Detail Enhancement Module (DEM), employing Laplacian-pyramid-based enhancement. This decomposition is instrumental in accentuating features obscured by adverse weather. The outputs from DENet are subsequently fused with the feature maps from PEFM, and this combined data is then fed into the detection head.

The fusion of feature maps from DENet and PEFM is governed by a decision score $\mathcal{S}$, obtained from the WDM, culminating in a final feature map $\mathcal{F}$ for object detection. Alongside these modules, we implemented a new domain align loss designed to bridge the domain gap between clear weather (source) and adverse weather (target) conditions, ensuring more consistent performance across different environments.

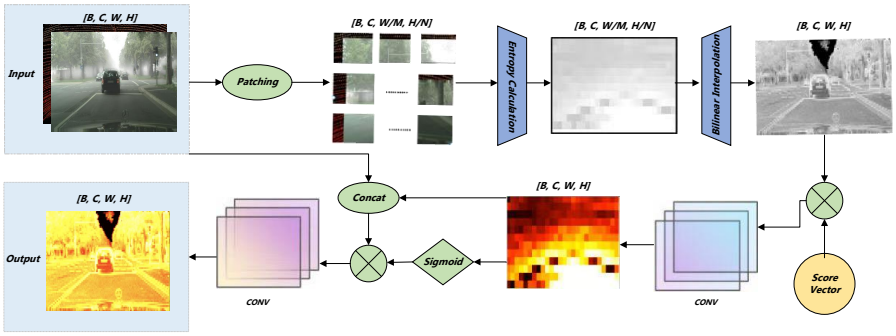

Fig. 3: This figure illustrates the workflow of the Patch Entropy Fusion Module (PEFM).

3.3 Weighted Decision Module

Given the variable performance of sensors across different adverse weather conditions—for instance, RGB cameras may excel in clear conditions but falter in fog, whereas LIDAR typically maintains effectiveness—the system incorporates a Weighted Decision Module (WDM). This module dynamically adjusts the weights assigned to various sensor features during the fusion process, thus tailoring the integration to adapt to the current environmental context.

As depicted in Fig. 2, inputs undergo initial processing through a 2D CNN block. This step enhances feature extraction utilizing the ReLU activation function [1]. Following this, a self-attention mechanism [34] combined with a linear and softmax layer [32] computes a score vector $\mathcal{S} = \{S_1, \ldots, S_j, \ldots, S_J\}$, where S_j corresponds to the weight assigned to the j^{th} sensor, with each score $S_j \in [0, 1]$. These scores critically determine the relevance of each sensor's feature map, thus influencing their contributions during the fusion process prior to reaching the detection head. This method ensures that each sensor's data is optimally utilized based on its effectiveness in specific weather conditions, enhancing the overall accuracy and robustness of object detection.

3.4 Patch Entropy Fusion Module

WDM can assign appropriate weights to data from different sensors, enhancing the integration of diverse sensory inputs. However, merely aligning features from distinct sensors is insufficient due to their inherent variability. For instance, RGB images are significantly affected by varying weather and lighting conditions, altering background appearance. Similarly, in LIDAR depth maps, points resulting from backscatter should not be indiscriminately aligned due to their distinct characteristics.

A noteworthy challenge arises when attempting to align features solely within the LIDAR or image data streams. Such an approach tends to bias the alignment

towards foreground regions, inadvertently diminishing the background representation. According to research highlighted in [3], leveraging local measurement entropy—a measure that assigns higher values to image regions with greater uncertainty, such as edges, corners, and foreground objects—can mitigate this issue. By multiplying deep feature representations with entropy values, this technique effectively diminishes background prominence while accentuating foreground features, thus enhancing object detection in complex visual scenes. However, this entropy-based approach has its limitations. Specifically, when feature maps from the LIDAR data stream are used as inputs for the detection algorithm, the LIDAR's entropy disproportionately influences the weighting of fused features. This can skew the detection model towards LIDAR-derived features, potentially overlooking valuable information from other sensory inputs.

To mitigate these challenges, our model refines the entropy scaling mechanism utilized in our baseline. Contrary to the strategy in [3], which relies on features from the LIDAR branch, our methodology projects all sensor data into the image plane and employs a modified entropy fusion scheme that we use entropy and the score vector to eliminate the noise information caused by the adverse weather to get domain-invariant foreground features. We calculate the entropy feature for each sensor as defined in Eq. 1:

$$\rho_j = f\left(\left(\sum_{m,n}^{w,h}\sum_{i=0}^{255}\sum_{t=1}^{3} p_{i,t}^{m,n}\log p_{i,t}^{m,n}\right)\right), \tag{1}$$

where w and h represent the width and height of the sensor stream, $f(.)$ denotes bilinear interpolation, and ρ_j is the entropy map for the j^{th} sensor. The entropy for each channel is calculated according to Eq. 2, considering an 8-bit input stream I with pixel values i ranging from 0 to 255 and t is the channel number:

$$p_{i,t}^{m,n} = \frac{1}{MN}\sum_{v,b}^{M,N}\varphi(I(m+v, n+b) - i). \tag{2}$$

Each sensor stream is partitioned into patches of size $M \times N$. Using the entropy map derived for each sensor, fusion is conducted following the weight scores from the WDM as per Eq. 3:

$$\rho = \begin{bmatrix} \frac{1}{J}\sum S_j \times \rho_{1,1,j} & \cdots & \frac{1}{J}\sum S_j \times \rho_{1,2,j} \\ \cdots & \cdots & \cdots \\ \frac{1}{J}\sum S_j \times \rho_{1,h,j} & \cdots & \frac{1}{J}\sum S_j \times \rho_{w,h,j} \end{bmatrix}. \tag{3}$$

The fused entropy map is subsequently processed by a 2D convolution layer and passed through a sigmoid function as shown in Fig. 3 to produce the final output from the PEFM as detailed in Eq. 4:

$$\mathcal{F}_p = \sigma\left(\mathrm{CNN}\left(\mho\left(I, \rho * I\right)\right)\right), \tag{4}$$

where σ symbolizes the sigmoid activation, CNN represents the 2D convolution operation, and $\mho$ signifies the element-wise multiplication followed by concatenation with the original input I.

3.5 Fusion and Detection

We adopt the DENet [29] branch, which decomposes the input data into high-frequency and low-frequency components using a Laplacian pyramid to produce the final feature map, denoted as $\mathcal{F}_d$. The set $\mathcal{F}_d = \mathcal{F}_{d,1}, \ldots, \mathcal{F}_{d,j}, \ldots, \mathcal{F}_{d,J}$ represents the feature maps obtained from the DENet for each sensor, where $\mathcal{F}_{d,j}$ corresponds to the feature map for the j^{th} sensor.

These feature maps from the DENet branch are then fused with those from the PEFN branch, guided by a score vector derived from the WDM. The fusion process is mathematically represented by the following equation:

$$\mathcal{F} = \mathcal{F}_p + \sum_j S_j \times \mathcal{F}_{d,j}, \tag{5}$$

Equation 5 defines the process for fusing the feature maps. Here, $\mathcal{S}_j \in \mathcal{S}$ represents the weight score assigned to the j^{th} sensor, indicating the confidence level in the sensor's data under the prevailing environmental conditions. $\mathcal{F}_p$ is the feature map derived from the PEFM, and $\mathcal{F}_{d,j}$ represents the feature maps from the DENet for each sensor. The fused feature map, $\mathcal{F}$, amalgamates these inputs, modulated by the WDM, into a comprehensive representation that enhances detection capabilities.

After the fusion process, $\mathcal{F}$ is fed into the detection head. This component can vary in architecture, including options such as a one-stage detector, a two-stage detector, or a transformer-based structure. Each type of detection head processes the input feature map $\mathcal{F}$ differently but computes two key types of outputs: the regression loss $\mathcal{L}_r$, which is responsible for bounding box prediction, and the classification loss $\mathcal{L}_c$, which classifies the detected objects within those boxes. These outputs are crucial for accurately locating and identifying objects within the input data, ultimately determining the effectiveness of object detection under varying environmental conditions.

3.6 Domain Align Loss

Training models directly on datasets collected under adverse weather conditions may lead to suboptimal performance due to sensor noise exacerbated by such environments. Drawing inspiration from domain adversarial training techniques [13], our approach processes source data from clear weather and target images from adverse weather simultaneously, aiming to align the corresponding feature distributions. We introduce a novel loss function aimed at mitigating the domain gap, thereby facilitating better alignment between distinct domains as formalized in Eq. 6:

$$\mathcal{L} = \mathcal{L}_d + \mathcal{L}_r + \mathcal{L}_c + \mathcal{L}_q, \tag{6}$$

where $\mathcal{L}_r$ and $\mathcal{L}_c$ represent the regression and classification losses from the detection head, respectively. The term $\mathcal{L}_d$ denotes the domain classification loss inspired by the work of [23]. This approach innovatively employs an adversarial gradient reversal layer (AdvGRL) that leverages negative gradients during

backpropagation. The aim is to confound the domain classifier, compelling it to generate features invariant to the domain. Subsequently, the output from the AdvGRL is fed into the domain classifier. This process is designed to facilitate global alignment, enhancing domain classification accuracy, as illustrated in Fig 2. The calculation of $\mathcal{L}_d$ is based on cross-entropy, as shown in Eq. 7:

$$\mathcal{L}_d = -\sum Y \log \hat{Y} + (1 - Y) \log(1 - \hat{Y}), \tag{7}$$

where $Y \in 0, 1$ corresponds to the ground-truth domain label, and $\hat{Y}$ is the predicted label from the domain classifier.

Furthermore, the loss function L_q, incorporates a feature metric constraint to effectively narrow the domain gap between the source and target domains. This distance metric is chosen for its ability to capture the maximum discrepancy between feature distributions, thereby providing a robust measure for domain adaptation. The formulation of L_q is as follows:

$$\mathcal{L}_q = \max\left(\sqrt{\sum(\mathcal{F}_{ds} - \mathcal{F}_{dt})^2}, 0\right), \tag{8}$$

this addition is predicated on the observation of source feature maps $\mathcal{F}_{ds}$ and target feature maps $\mathcal{F}_{dt}$.

4 Experiment

In this section, we will discuss the outcomes of our experimental evaluation. First, we will detail the implementation specifics, which include the configuration settings and the computational environment used to conduct the experiments. This will be followed by a comprehensive overview of the datasets employed and the baseline models against which our results are benchmarked. Subsequently, we will conduct an extensive ablation study to analyze the impact of each component within our model on its overall performance. Finally, we will compare our experimental results with those of the baseline models.

4.1 Experiment Setup

For our training parameters, we initiated with a learning rate of 0.0001 and set the weight decay at 0.0005. The Non-Maximum Suppression (NMS) threshold was established at 0.5 to balance detection precision and recall. We utilized a batch size of 8 and selected the Adam optimizer [21] paired with a Cosine learning rate scheduler to optimize and adapt the learning rate throughout the training duration. Additionally, the momentum was set at 0.9. This training regimen spans 300 epochs, incorporating an early stopping mechanism to mitigate the risk of overfitting. To ensure a thorough and equitable evaluation, our architecture includes several detection heads. This design choice caters to comparative analysis with a broad spectrum of baseline models, some of which are based on

(a) ExDark dataset

(b) Cityscapes dataset

(c) Dense dataset

Fig. 4: This figure presents the detection results for three distinct datasets. Subfigure (a) displays the results from the ExDark dataset. Subfigure (b) illustrates the results from the Cityscapes dataset. Subfigure (c) depicts the results from the Dense dataset

YOLOv3 [30], while others leverage Fast R-CNN [15]. Our model dynamically selects the appropriate detection head to align with the evaluated baseline, facilitating direct and fair comparisons. The implementation of our model leverages the PyTorch framework. All experiments are executed on a high-performance computing setup consisting of a single NVIDIA GeForce RTX 4090 GPU and a 12th Generation Intel(R) Core(TM) i9-12900k CPU.

4.2 Datasets and Baseline

Our model is assessed under challenging visibility conditions, specifically low light and foggy weather, to validate its robustness and adaptability. The datasets selected for evaluation cater to these specific environmental conditions:

– **Low-Light Condition**: We utilize the Exclusively Dark (ExDark) dataset [25], which comprises 7,363 low-light images across 12 object categories. This unimodal dataset is designed to benchmark performance in low-visibility conditions.

(a) Detection results with all modules enabled (b) Detection results with all modules disabled

Fig. 5: This figure presents a comparative analysis of the results obtained with and without all modules across three distinct datasets.

– **Foggy Weather**: Evaluation under foggy conditions uses two datasets. The first is the Cityscapes-to-Foggy Cityscapes [8], a real-world, task-driven testing set that simulates foggy weather scenarios. We evaluate Cityscapes with AP for each class, including car, bicycle, bus, person and rider, and we use mAP for the total. The second is the Dense Fog Day dataset [3], further challenging the model with dense fog conditions. We evaluate our approach in three classes: Passenger Car (Car), Pedestrian (Ped.), and Ridable Vehicle (RV) with KITTI [14] framework. We split the Dense dataset as the same as [11]

For the ExDark dataset, our model is benchmarked against leading methods optimized for low-light conditions, including MS-DAYOLO [18], DAYOLO [38], MAET [9], and DE-YOLO [29]. These selections are based on their demonstrated effectiveness in similar visibility challenges.

For the Cityscapes dataset, the comparison extends to recent domain adaptation methods that have shown promise in foggy conditions. These include MTOR [5], DA-Faster [7], GPA [37], UaDAN [17], and DA-Detect [23], providing a comprehensive overview of our model's performance relative to state-of-the-art techniques in domain adaptation.

For the Dense dataset, we compare our approach against advanced domain adaptation and image translation methodologies, such as ADDA [33], CyCADA [19], SMTDA [11], and Dense [3]. This comparison highlights our model's capability to handle extreme visibility conditions through effective domain adaptation.

4.3 Performance Analysis

Our comprehensive evaluation across diverse environmental conditions demonstrates the robustness and effectiveness of our proposed method. We performed extensive comparisons on three different datasets, each addressing unique challenges posed by low-light, urban, and foggy conditions. We present the visualization of the testing detection results in Fig. 5. Below, we discuss the results from these evaluations, which are summarized in Tables 1, 2, and 3.

Cityscapes Dataset: The detailed performance metrics are presented in Table 1, where our model notably excels, achieving the highest mAP of 49.1%. Our method shows performance in detecting cars and persons, with AP scores of

Table 1: AP for each class and overall mAP with comparison methods on the Cityscapes.

Method	Modality	Car ↑(%)	Bicycle ↑(%)	Bus ↑(%)	Person ↑(%)	Rider ↑(%)	mAP ↑(%)
MTOR [5]	RGB	44.0	35.6	38.6	30.6	41.4	35.1
DA-Faster [7]	RGB	53.0	39.0	49.8	35.7	45.2	41.0
GPA [37]	RGB	54.1	38.7	45.7	32.9	46.7	39.5
UaDAN [17]	RGB	53.6	38.9	49.4	36.5	36.1	41.1
DA-Detect [23]	RGB	54.3	39.1	**51.2**	36.5	46.7	42.3
Ours	RGB	**67.8**	**40.2**	42.4	**45.9**	**49.7**	**49.1**

Table 2: Comparisons on ExDark.

Method	Modality	AP50 ↑(%)
MS-DAYOLO [18]	RGB	44.3
DAYOLO [38]	RGB	44.6
MAET * [9]	RGB	47.1
DE-YOLO [29]	RGB	51.5
Ours	RGB	**53.8**

* The AP50 is taken from [29].

Table 3: Comparison on Dense.

Method	Modality	Car ↑(%)	Ped ↑(%)	RV ↑(%)
ADDA [33]	Unimodal	83.5	38.6	2.5
CyCADA [19]	Unimodal	82.8	33.5	1.3
SMTDA [11]	Multimodal	85.9	45.0	4.6
Dense [3]	Multimodal	84.0	38.2	5.8
Ours	Multimodal	**87.8**	**48.2**	**6.0**

67.8% and 45.9%, respectively. These results are higher than competing methods like DA-Detect [23], which previously led with an mAP of 42.3%.

ExDark Dataset: As shown in Table 2, our method achieves a superior average precision (AP50) of 53.8%, which is higher than that of other well-established methods. Our model outperforms the next best method, DE-YOLO [29], by over two percentage points.

Dense Dataset: As Table 3, our model outstrips all other domain adaptation and image translation techniques, achieving the highest scores across all categories, with a standout AP of 87.8% for cars and 48.2% for pedestrians.

4.4 Ablation Study

In this section, we discuss the ablation study results for the WDM, PEFM, and Domain Alignment Loss (DALOSS) across the three datasets, and we also compare the detection results obtained with all modules enabled against those without any modules, as shown in Fig 5.

ExDark Dataset: Primarily utilizing RGB data, this unimodal dataset allows us to evaluate the effectiveness of PEFM and DALOSS. With PEFM and DALOSS enabled, the model reaches an average precision (AP) of 53.8%. Removal of DALOSS decreases AP to 52.2%, while disabling PEFM further reduces it to 51.7%.

Dense Dataset: This multimodal dataset, which includes categories such as Cars, Pedestrians, and Road Vehicles (RV), demonstrates the benefits of a fully integrated approach. When WDM, PEFM, and DALOSS are all active, the model achieves its highest performance: 87.8% AP for Cars, 48.2% for Pedestrians, and 6.0% for RVs.

Table 4: Ablation experiment results of ap for ExDark.

Dataset	WDM	PEFM	DALOSS	AP ↑(%)
ExDark	–	✓	✓	**53.8**
	–	✓	×	52.2
	–	×	✓	47.7
	–	×	×	45.5

Table 5: Ablation experiment results of ap for Dense.

Dataset	WDM	PEFM	DALOSS	Car	Ped	RV
DENSE	✓	✓	✓	**87.8**	**48.2**	**6.0**
	✓	✓	×	86.2	45.5	4.2
	×	✓	✓	85.8	46.2	5.3
	✓	×	✓	83.8	45.6	3.5

Table 6: Ablation experiment results of mAP for CITYSCAPES.

Dataset	WDM	PEFM	DALOSS	mAP ↑(%)
CITYSCAPES	–	✓	✓	**49.1**
	–	✓	×	46.8
	–	×	✓	41.5

Cityscapes Dataset: This urban scene dataset emphasizes the value of integrating comprehensive feature fusion and domain alignment strategies. Employing both PEFM and DALOSS achieves the highest mAP of 49.1%. Omitting DALOSS results in a reduced mAP of 46.8%, and the absence of PEFM leads to a pronounced decrease to 41.5%. The findings highlight the critical role of DALOSS in enhancing domain robustness, which is crucial for adapting to the diverse urban environments characteristic of the Cityscapes dataset.

5 Conclusion

In this paper, we introduce a decision and entropy-based object detection model designed to enhance the robustness of object detection tasks in various adverse weather conditions. Our innovative approach integrates a data augmentation scheme that effectively simulates diverse weather scenarios and an advanced entropy fusion strategy for extracting complementary information across different modalities. Moreover, our model incorporates a weighted decision module that strategically allocates weights to different sensory inputs based on the prevailing weather conditions if multimodal are used. Additionally, we employ a domain alignment loss function that effectively minimizes the domain discrepancies, thereby improving model generalizability across different environmental settings. Collectively, these contributions not only advance the state-of-the-art in weather-robust object detection but also pave the way for future research in domain-adaptive computer vision systems.

Acknowledgments. The works were jointly supported by the Suzhou Science and Technology Development Planning Programme (Grant No.ZXL2023171) and XJTLU Research Development Fund (RDF-22-01-129)

References

1. Agarap, A.F.: Deep learning using rectified linear units. arXiv preprint arXiv:1803.08375 (2018)
2. Asvadi, A., Garrote, L., Premebida, C., Peixoto, P., Nunes, U.J.: Multimodal vehicle detection: fusing 3d-lidar and color camera data. Pattern Recognition Letters **115**, 20–29 (2018)
3. Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., Heide, F.: 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. pp. 11682–11692 (2020)
4. Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., et al.: Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 4243–4250. IEEE (2018)
5. Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T.: Exploring object relation in mean teacher for cross-domain detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11457–11466 (2019)
6. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1907–1915 (2017)
7. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
9. Cui, Z., Qi, G.J., Gu, L., You, S., Zhang, Z., Harada, T.: Multitask aet with orthogonal tangent regularity for dark object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2553–2562 (2021)
10. Du, X., Ang, M.H., Rus, D.: Car detection for autonomous vehicle: Lidar and vision fusion approach through deep learning framework. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 749–754. IEEE (2017)
11. Eskandar, G., Marsden, R.A., Pandiyan, P., Döbler, M., Guirguis, K., Yang, B.: An unsupervised domain adaptive approach for multimodal 2d object detection in adverse weather conditions. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10865–10872. IEEE (2022)
12. Feng, D., Harakeh, A., Waslander, S.L., Dietmayer, K.: A review and comparative study on probabilistic object detection in autonomous driving. IEEE Transactions on Intelligent Transportation Systems **23**(8), 9961–9980 (2021)
13. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
14. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
15. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)

16. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. Journal of field robotics **37**(3), 362–386 (2020)
17. Guan, D., Huang, J., Xiao, A., Lu, S., Cao, Y.: Uncertainty-aware unsupervised domain adaptation in object detection. IEEE Transactions on Multimedia **24**, 2502–2514 (2021)
18. Hnewa, M., Radha, H.: Multiscale domain adaptive yolo for cross-domain object detection. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 3323–3327. IEEE (2021)
19. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. pp. 1989–1998. Pmlr (2018)
20. Huang, S.C., Le, T.H., Jaw, D.W.: Dsnet: Joint semantic learning for object detection in inclement weather conditions. IEEE transactions on pattern analysis and machine intelligence **43**(8), 2623–2633 (2020)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Langer, F., Milioto, A., Haag, A., Behley, J., Stachniss, C.: Domain transfer for semantic segmentation of lidar data using deep neural networks. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8263–8270. IEEE (2020)
23. Li, J., Xu, R., Ma, J., Zou, Q., Ma, J., Yu, H.: Domain adaptive object detection for autonomous driving under foggy weather. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 612–622 (2023)
24. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 21–37. Springer (2016)
25. Loh, Y.P., Chan, C.S.: Getting to know low-light images with the exclusively dark dataset. Computer Vision and Image Understanding **178**, 30–42 (2019)
26. Mees, O., Eitel, A., Burgard, W.: Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 151–156. IEEE (2016)
27. Paz, D., Zhang, H., Li, Q., Xiang, H., Christensen, H.I.: Probabilistic semantic mapping for urban autonomous driving applications. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2059–2064. IEEE (2020)
28. Pfeuffer, A., Dietmayer, K.: Optimal sensor data fusion architecture for object detection in adverse weather conditions. In: 2018 21st International Conference on Information Fusion (FUSION). pp. 1–8. IEEE (2018)
29. Qin, Q., Chang, K., Huang, M., Li, G.: Denet: detection-driven enhancement network for object detection under adverse weather conditions. In: Proceedings of the Asian Conference on Computer Vision. pp. 2813–2829 (2022)
30. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
31. Schneider, L., Jasch, M., Fröhlich, B., Weber, T., Franke, U., Pollefeys, M., Rätsch, M.: Multimodal neural networks: Rgb-d for semantic segmentation and object detection. In: Image Analysis: 20th Scandinavian Conference, SCIA 2017, Tromsø, Norway, June 12–14, 2017, Proceedings, Part I 20. pp. 98–109. Springer (2017)
32. Tang, J., Deng, C., Huang, G.B.: Extreme learning machine for multilayer perceptron. IEEE transactions on neural networks and learning systems **27**(4), 809–821 (2015)

33. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
35. Wang, H., Xu, Y., He, Y., Cai, Y., Chen, L., Li, Y., Sotelo, M.A., Li, Z.: Yolov5-fog: A multiobjective visual detection algorithm for fog driving scenes based on improved yolov5. IEEE Transactions on Instrumentation and Measurement **71**, 1–12 (2022)
36. Wang, Z., Jia, K.: Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1742–1749. IEEE (2019)
37. Xu, M., Wang, H., Ni, B., Tian, Q., Zhang, W.: Cross-domain detection via graph-induced prototype alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12355–12364 (2020)
38. Zhang, S., Tuo, H., Hu, J., Jing, Z.: Domain adaptive yolo for one-stage cross-domain detection. In: Asian conference on machine learning. pp. 785–797. PMLR (2021)
39. Zhu, Y., Wang, T., Fu, X., Yang, X., Guo, X., Dai, J., Qiao, Y., Hu, X.: Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21747–21758 (2023)