ISSN: 2414-1895 DOI: 10.6919/ICJE.202607_12(7).0010

YOLO-LMC: An Enhanced Framework for Robust Object Detection

Gang Lin[1], Senming Li[1], Zhefeng Yu[1], Yongnan Leng[1], Hongjian Lu[1], Wei Zhang[1], Yongwei Ma[1], Xiaoman Wang[1], and Juncheng Jiang[2, *]

[1] Zhejiang Zheneng Jinhua Gas Turbine Power Generation Co., Ltd, Jinhua 321000, China

[2] Huzhou Institute of Zhejiang University, Huzhou 313000, China

Abstract

Object detection in hazy scenes holds significant real-world implications, particularly in areas such as intelligent transportation and environmental monitoring. However, image quality degradation caused by haze inherently weakens the feature extraction capabilities of neural network models, leading to reduced accuracy. To address this challenge, this paper proposes YOLO-LMC (YOLO with Latent-space Metric Constraints), a novel architecture that integrates metric learning within the latent space optimization of object detection. Specifically, a positive sample assignment strategy is designed to filter high-confidence features within the latent space. This strategy, coupled with a cosine distance metric loss function based on class centers, effectively compels intra-class samples to form compact distributions in the latent space while ensuring distinct inter-class separability. Experimen-tal validation on the Foggy-Cityscapes dataset demonstrates that the proposed method significantly outperforms other object detection enhancement approaches, achieving an impressive at least 8.6% improvement in mAP.

Keywords

Hazy Scenes; Object Detection; YOLO-LMC; YOLO-v8; Latent Space Metric Learning; Positive Sample Assignment.

1. Introduction

Object detection is a fundamental task in intelligent visual perception and has been widely applied in intelligent transportation, environmental monitoring, and industrial inspection. In practical scenarios such as gas turbine power plants, it also supports equipment inspection, operational monitoring, and abnormal event detection. However, haze-like interference, including fog, water vapor, and exhaust plumes, often degrades image quality and weakens feature representation, leading to reduced detection accuracy. Although existing methods based on image enhancement, dehazing, or domain adaptation can alleviate this issue to some extent, they often rely on restoration quality or lack explicit constraints on latent feature separability, resulting in limited robustness under degraded conditions [1-5].

However, the performance of CNN-based models generally depends heavily on high-quality, large-scale datasets. Although existing object detection datasets provide abundant image samples and high-quality annotations, enabling satisfactory performance under standard conditions, these datasets are mostly collected in environments with stable illumination, clear backgrounds, and high imaging quality [6]. Studies have shown [7] that haze significantly degrades image contrast and target texture information, interferes with the extraction of critical features by CNNs, and thus leads to a substantial decline in detection performance. Although retraining models with newly collected images can improve their adaptability to hazy conditions, this process usually requires considerable human effort, material resources, and time [8].

Existing methods for object detection under degraded conditions mainly focus on image enhancement, dehazing, or domain adaptation. However, they often depend on restoration quality or fail to explicitly enforce latent feature separability, resulting in limited robustness and accuracy. To address the challenges of object detection under hazy conditions, existing research has mainly followed two technical routes. Current methods [9] can generally be divided into two categories: one focuses on image enhancement or dehazing before detection, while the other employs feature alignment or domain adaptation strategies. The first category typically integrates a dehazing module into the detection framework in order to restore clearer scene information before feeding images into the detector. For example, Chengyang Li et al. [7] proposed BAD-Net, which jointly connects dehazing and detection modules in an end-to-end manner. This architecture adopts a dual-branch structure with an attention fusion mechanism to collaboratively model hazy and dehazed features, while also introducing a self-supervised haze-robust loss function to improve robustness under different haze densities. Similarly, Teena Sharma et al. [10] proposed FR-HDNet, a Faster R-CNN-based haze detection network that identifies and removes hazy regions, producing enhanced images better suited for subsequent visual tasks.Cai et al. [11] proposed DehazeNet, an end-to-end CNN for medium transmission estimation with strong dehazing capability. However, dehazing-before-detection methods are highly dependent on dehazing quality, and the resulting artifacts or distortions may degrade downstream detection performance. Moreover, they often do not directly enhance haze-robust feature learning in the latent space.

Domain adaptation has been extensively studied for object detection in degraded and cross-domain scenarios. Zhu et al. [12] improved detection via selective cross-domain alignment of salient local regions, while Zhao et al. [13] enhanced cross-domain performance by jointly aligning feature space, semantic content, and style information. Hnewa et al. [14] proposed MS-DAYOLO to improve domain invariance at multiple detection scales under adverse weather conditions. Saito et al. [15] combined strong local alignment with weak global alignment for unsupervised detector adaptation, and Wu et al. [16] further introduced THNet, which integrates instance-, pixel-, and image-level alignment for robust cross-domain detection. Nevertheless, these methods may still struggle in complex real-world scenes, particularly for small objects, and often lack explicit constraints on latent feature discriminability.

In light of the aforementioned issues, This study proposes a novel detection framework based on latent space feature metric constraints, termed YOLO-LMC (YOLO with Latent-space Metric Constraints).This approach aims to enhance the robustness of extracted features against cross-domain targets. Its effectiveness is rigorously verified through experiments on the Foggy-Cityscapes dataset, utilizing depth information. The main contributions are summarized as follows:
A masked task alignment learning (MTAL) method tailored for the latent space is designed. Positive samples in the latent space are selected based on the matching relationship between labels and predictions, and the corresponding category relationships are clarified.

A category-center-based metric learning architecture based on cosine similarity is proposed. By minimizing the distance between samples of the same category and the category center, and maximizing the distance between different category centers, the inter-class separability in the latent space is significantly enhanced.

2. Methodology

2.1 YOLO-LMC

The proposed algorithm framework is shown in Figure 1and is based on improve- ments to YOLOv8 [17]. The network consists of a backbone network, a Feature Pyramid Network (FPN) [18], a Latent Space Metric (LMC)[19] Module, and a detection head.The technical implementation process can be broken down into the following steps:

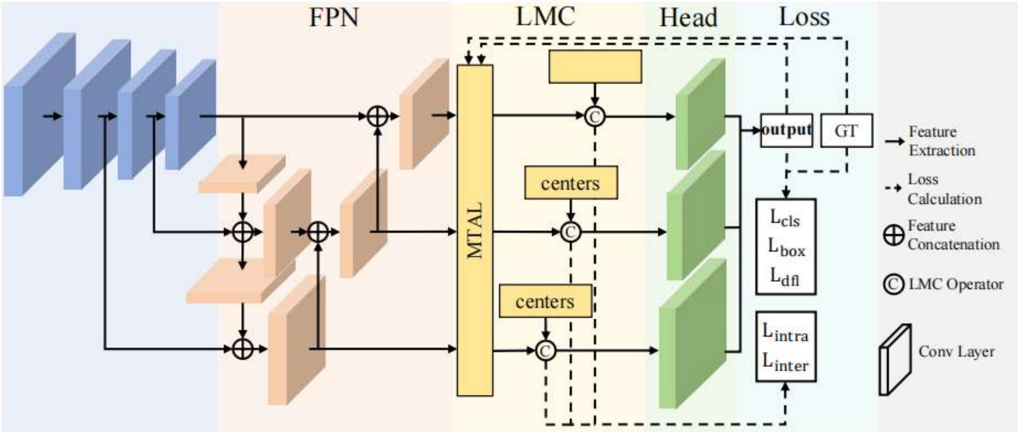

Figure 1. Framework of the YOLO-LMC algorithm

Firstly, the input image is processed through the DarkNet-CSP [20] backbone network for hierarchical convolutional operations,generating multi-scale high-dimensional semantic feature maps.Deep features carry global semantic information, while shallow features preserve fine-grained texture details.

Secondly, the FPN is used to perform top-down hierarchical feature fusion, constructing a latent space representation with cross-scale semantic associations. Next,a metric optimization mechanism is designed in the latent space to calculate similarity metrics between samples of the same class and between samples of different classes. This results in intra-class loss L_{intra} and inter-class loss L_{inter}. Adaptive learning optimization is employed to achieve compactness in the intra-class feature distribution.

Finally, the optimized latent features are processed by the detection head for class distinction and bounding L_{box} regression, yielding the final prediction results. These prediction results are used to compute the classification loss L_{cls}, regression loss L box, and distribution focal loss L_{dfl} with the labels.

2.2 Positive Sample Selection in the Latent Space

When calculating the loss in the latent space, since the output of the latent space is a high-density feature map, and there inevitably exist effective features describing targets and ineffective features describing backgrounds, calculating the loss over the entire map will inevitably introduce noise from too many invalid features, weakening the model’s learning ability. Therefore, it is necessary to screen the feature maps and select positive samples to participate in the loss calculation.

To this end, the Masked Task Alignment Learning (MTAL) method is proposed. On the basis of Task Alignment Learning (TAL) [21,22], a dynamic threshold matrix is set to filter out predictions with low confidence, thereby enhancing the quality of positive sample selection. The specific steps are as follows:

(1) Preliminary screening: The GT (Ground Truth) bounding box is scaled to the corresponding feature map scale. If a sample point is within the scaled GT box, it is selected as a positive sample, and the rest are negative samples.

(2) Mixed score matrix calculation: The results of the preliminary screening in the first step may still contain some negative samples, such as those with low classification confidence in the prediction or large bounding box offsets. Further screening is needed. According to the prediction results, calculate the CIOU [23] value between each predicted box and the GT box, and combine the classification probability of the predicted box to obtain the mixed score matrix T. The physical meaning of T is the comprehensive prediction probability of the sample point on each GT box. The specific calculation formulas are shown in Equations 1, 2, 3, and 4:

$$CIOU = IOU - \frac{d^2}{c^2} - \alpha v \tag{1}$$

$$IOU = \frac{B_P \cap B_G}{B_P \cup B_G} \tag{2}$$

$$\alpha = \frac{\upsilon}{(1 - IOU) + \upsilon} \tag{3}$$

$$\upsilon = \frac{4(\arctan\frac{\omega_G}{h_G} - \arctan\frac{\omega_p}{h_p})^2}{\pi^2} \tag{4}$$

where d is the distance between the center points of the predicted box and the ground truth box, c is the diagonal distance of the smallest enclosing rectangle, v is a correction factor used to further adjust the loss function, B_P and B_G are the predicted box and the ground truth box respectively, IOU is the ratio of the intersection area to the union area of the predicted box and the ground truth box, and w_G, h_G, w_P, h_P are the widths and heights of the ground truth box and the predicted box respectively. The CIOU value and the classification score Score obtained from the prediction are substituted into the following formula:

$$T = Score^{\alpha} \cdot CIOU^{\beta} \tag{5}$$

where α and β are hyperparameters used to balance the convergence speed of classification accuracy and localization accuracy. Sort T to obtain the top K positive samples, and the rest are considered as negative samples. As shown in Equation 6:

$$T_{topK} = \left\{ T_{n,p} \middle| p \in \arg \max_{p' \in \{1,\ldots,K\}}^{k} T_{n,p'} \right\} \tag{6}$$

(3) Constructing a dynamic threshold matrix to filter low-quality samples: There are predictions with low scores in the mixed score matrix. It is necessary to design an adaptive threshold for each GT box based on the comprehensive prediction situation to filter out low-quality samples. Specifically, for the top K prediction results of a certain GT box, if its mean value is high, it indicates that the GT box is of good quality and easy to distinguish. The threshold should be correspondingly increased to avoid too many positive samples being selected, which would cause the model to overly focus on easily distinguishable samples. On the contrary, if the mean value is low, it indicates that the GT box is of poor quality. The threshold should be correspondingly decreased to include more samples for optimization.For a certain GT box, if the standard deviation of its top K prediction results is high, it means that the model's predictions for this GT box fluctuate greatly. The threshold should be increased to avoid the introduction of noise. Conversely, if the standard deviation is low, it means that the model's predictions for this GT box are more consistent. The threshold can be appropriately decreased to include more samples to help convergence. Therefore, the dynamic threshold matrix is defined as follows:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202607_12(7).0010

$$\mu_n = \frac{1}{N_G}\sum_{n}^{N} T_{topK_{n,p'}} \tag{7}$$

$$\sigma_n = \sqrt{\frac{1}{p}\sum_{p=1}^{p}(T_{topK_{n,p}} - \mu_n)^2} \tag{8}$$

$$M_n = \mu_n + \sigma_n \tag{9}$$

where μ is the mean of the top K scores T in the mixed score matrix, σ is the variance, N_G is the number of GT boxes, P is the total number of samples, and M is the dynamic threshold matrix. Use the dynamic threshold matrix M to filter the mixed score matrix T, retaining the part where $T > M$, as shown in Equation 10:

$$T_{n,p} = \begin{cases} T_{n,p} & T_{n,p} \geq M_n \\ 0 & others \end{cases} \tag{10}$$

(4)Eliminate duplicate assignments: Exclude cases where a sample is assigned to multiple GT boxes. Traverse the number of GT boxes assigned to each sample. If it is greater than 1, it indicates a duplicate assignment. For samples with duplicate assignments, only the one with the highest CIOU is retained. In this way, positive samples can be screened out, with each positive sample corresponding to a GT, thereby establishing the correspondence between positive samples in the latent space and GT categories.

The pseudocode of the proposed MTAL is shown in Algorithm 1, where B represents the Batch Size, N_G represents the number of GT boxes, N_P represents the number of predicted boxes, P represents the number of samples, and D represents the number of channels in the latent space feature map.

2.3 Clustering Center Metric Based on Cosine Similarity

After the positive samples have been screened, it is necessary to optimize the feature space to achieve the following goals: enhance the feature aggregation degree of positive samples of the same class, expand the feature separability of samples from different classes, and thereby construct a class-internal feature distribution structure with strong discrim-inability. To achieve this goal, cosine similarity is adopted as the core metric criterion, based on the following considerations:

The cosine similarity function exhibits a "gentle-steep-gentle" nonlinear trend in the angle space [0, π]. When the angle between feature vectors is close to 0 (for samples of the same class), the cosine similarity is less sensitive to angle changes, which helps mainta in the stability of intra-class features. When the angle is close to *π/2* (the potential decision- boundary region), the cosine similarity is highly sensitive to small angle deviations, with a large gradient change. This characteristic forces the model to strengthen the learning of feature boundaries for difficult-to-distinguish samples. When the angle is close to *π* (for samples of different classes), the similarity enters a low-sensitivity zone again, avoiding overly strong constraints on clearly separated samples of different classes.

Then, a parametric multimodal clustering center learning strategy is proposed. The core design is to define a set of learnable clustering centers for each category. These clustering centers are jointly optimized with the model through gradient descent, enabling them to adaptively fit the dynamic distribution characteristics of categories in the latent space. At the same time, the clustering centers provide anchor points for aggregation in the latent space.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202607_12(7).0010

Targets are distributed differently in space and have different poses. When expressed in the latent space, predictions for the same class may have different modalities. To enhance the model's ability for bounding box regression, K sub-clustering centers are preset for each category, representing different modality prototypes of the category and improving the fitting ability of the intra-class feature structure. The specific algorithm process is as follows:

Define clustering centers: Randomly initialize clustering centers $C_K^{N_{cls}} \in R^{N_{cls},K}$, where B is the Batch_size, K is the number of sub-clustering centers, and N_{cls} is the number of classes;

(2)Intra-class similarity loss calculation:Calculate the cosine similarity between positive sample features and their corresponding sub-clustering centers, and the intra-class loss, as shown in Equations 11 and 12:

$$D_{\mathrm{intra}_i} = \frac{1}{N_{pos}K}\frac{XC_i}{\|X\|_2\|C_i\|_2} \in R^{N_{pos},K} \tag{11}$$

$$L_{\mathrm{intra}} = \frac{1}{N_{cls}}\sum_{i=1}^{N_{cls}} D_{\mathrm{intra}_i} \tag{12}$$

where X represents the selected positive samples, C_i represents the sub-clustering centers of the i-th class, N_{pos} is the number of positive samples screened by MTAL, K is the number of sub-clustering centers, and N_{cls} is the number of classes.

(3) Inter-class similarity loss calculation:Calculate the cosine similarity between clustering centers of different classes and the inter-class loss, as shown in Equations 13 and 14:

$$D_{\mathrm{inter}_{i,j}} = \frac{C_i C_j^T}{\|C_i\|_2\|C_j\|_2} \tag{13}$$

$$L_{\mathrm{inter}} = \frac{1}{(N_{cls}-1)^2}\sum_{i=1}^{N_{cls}}\sum_{j=1,j\neq i}^{N_{cls}} D_{\mathrm{inter}_{i,j}} \tag{14}$$

(4) Overall metric loss calculation: After obtaining the intra-class and inter-class losses, calculate the overall metric loss according to Equation 15, which, together with the original classification loss L_{cls}, localization loss L_{box}, and distribution focal loss L_{dlf} of YOLOv8, participates in the back-propagation calculation. The margin represents the interval between inter-class and intra-class distances, which is used to enhance the model's discriminative ability.

$$L_{\mathrm{metric}} = L_{\mathrm{intra}} - L_{\mathrm{inter}} + margin \tag{15}$$

The pseudocode of the proposed clustering center metric strategy based on cosine similarity is shown in Algorithm 2, where B represents the Batch Size, N represents the number of classes, M represents the number of positive samples selected, and D represents the number of channels in the latent space feature map.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202607_12(7).0010

3. Experiments and Results

3.1 Dataset

To systematically evaluate the performance of YOLO-LMC in hazy conditions, the YOLO-LMC model was trained on the Cityscapes[24] dataset and validated on the Foggy-Cityscapes dataset [25].

The Cityscapes dataset is a widely recognized benchmark for urban scene understand-ing, particularly in the context of autonomous driving. It provides a large-scale collection of diverse street scenes captured from 50 different cities, encompassing varying weather conditions and times of day. The images maintain a high resolution of 1024x2048 pixels and cover over 30 categories of common urban elements. For most research applications of object detection, a consolidated set of 8 common categories, as detailed in Table 1, often suffices, with the dataset providing approximately 5000 finely annotated images and 60,000 instances.

The Foggy-Cityscapes is a synthetic dataset generated from the well-established Cityscapes dataset, specifically designed to simulate and analyze the performance of autonomous vehicles under foggy conditions. The dataset features three distinct levels of fog density (0.005, 0.01, and 0.02), as shown in Figure 2, corresponding to visibility ranges of 600, 300, and 150 meters, respectively, representing a spectrum from light mist to dense fog. Crucially, Foggy-Cityscapes retains the comprehensive annotation information present in the original Cityscapes dataset.

Table 1. Statistics of Cityscapes Classes and Instances.

Class	person	rider	car	truck	bus	train	mcycle	bicycle
Count	21413	2363	31822	582	483	194	888	4904

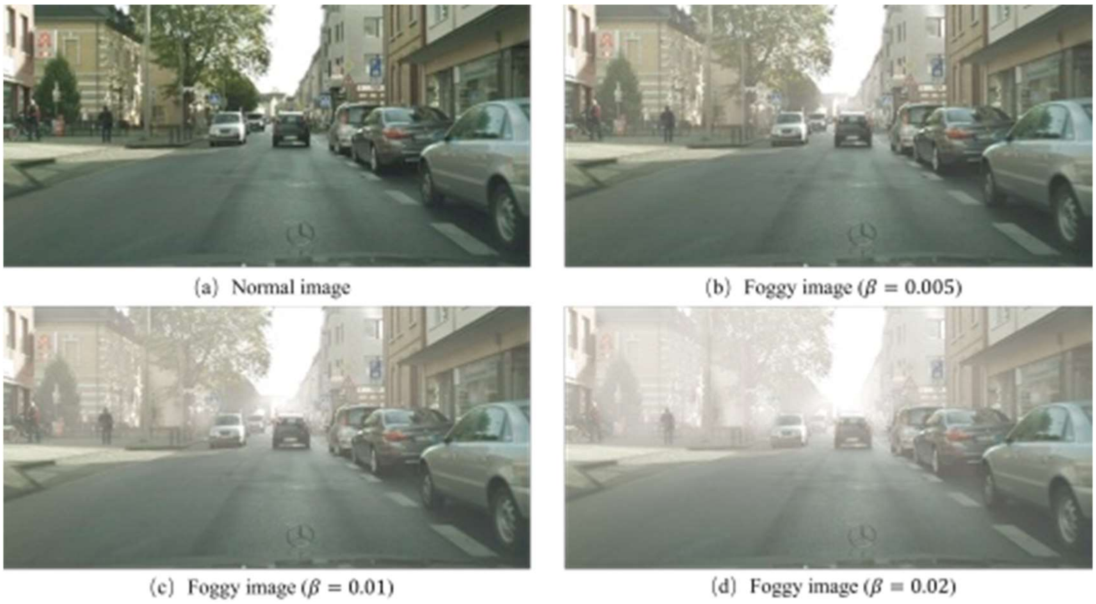

Figure 2. Example images from Foggy Cityscapes

3.2 Evaluation Metrics

The evaluation metrics selected are the average precision (AP) and the mean average precision (mAP). These metrics reflect the comprehensive accuracy of model localization and classification.

In addition to these, the Kullback-Leibler Divergence (KL Divergence) [26] is used to measure the distribution differences between predictions of different classes and between predictions of the same class in different scenarios. KL Divergence is a tool in information theory used to measure the difference between two probability distributions. For two discrete probability distributions P and Q, the KL Divergence is defined as shown in Equation 16, representing the additional amount of

 Volume 12 Issue 7, 2026
DOI: 10.6919/ICJE.202607_12(7).0010

information needed when approximating the true distribution P with distribution Q. And a larger value indicates a greater difference between the two distributions.

$$D_{KL}(P\|Q)=\int_{-\infty}^{\infty}P(x)\ln\frac{P(x)}{Q(x)}dx \tag{16}$$

In the experiments, since the exact distribution in the latent space is unknown, a data-driven method based on equal-width binning is used to calculate the KL Divergence. After MTAL allocation, the obtained positive sample set is $X\in R^{M,D}$, with corresponding labels $Y \in \{1,...,C\}^{M}$, where C is the number of classes, M is the number of positive samples, and D is the number of channels. The positive samples of the c-th class are denoted as $X^{(c)}\in R^{Mc,D}$, with Mc being the number of positive samples of the c-th class. Equal-width binning divides the maximum and minimum values of feature dimension $d \in \{1,...,D\}$ into K equal-width intervals (bins), with bin boundaries as shown in Equation 17, where min (·) and max (·) represent the minimum and maximum values of the current dimension, respectively.

$$b_k^{(c)}=\min(X_d^{(c)})+\frac{k}{K}\left[\max\left(X_d^{(c)}\right)-\min\left(X_d^{(c)}\right)\right],\forall k\in\{0,\dots,K\} \tag{17}$$

The probability mass function estimation after binning discretization is shown in Equation 18, where **I**(·) is the indicator function, taking 1 if the condition in parentheses is met and 0 otherwise, and ε is a smoothing coefficient, typically 10^{-8}:

$$P_d^{(c)}(k)=\frac{1}{M_c}\sum_{i=1}^{M_c}\mathrm{I}\left(X_{i,d}^{(c)}\in\left[b_{k-1}^{(c)},b_k^{(c)}\right)\right)+\varepsilon \tag{18}$$

The calculation method of KL Divergence for classes c*1* and *c2* is shown in Equation 19:

$$D_{KL}\left(p_{c1}\|p_{c2}\right)=\sum_{d=1}^{D}\sum_{k=1}^{K}p_d^{(c1)}(k)\log\frac{p_d^{(c1)}(k)}{p_d^{(c2)}(k)} \tag{19}$$

3.3 Analysis of Latent Space Distribution Differences

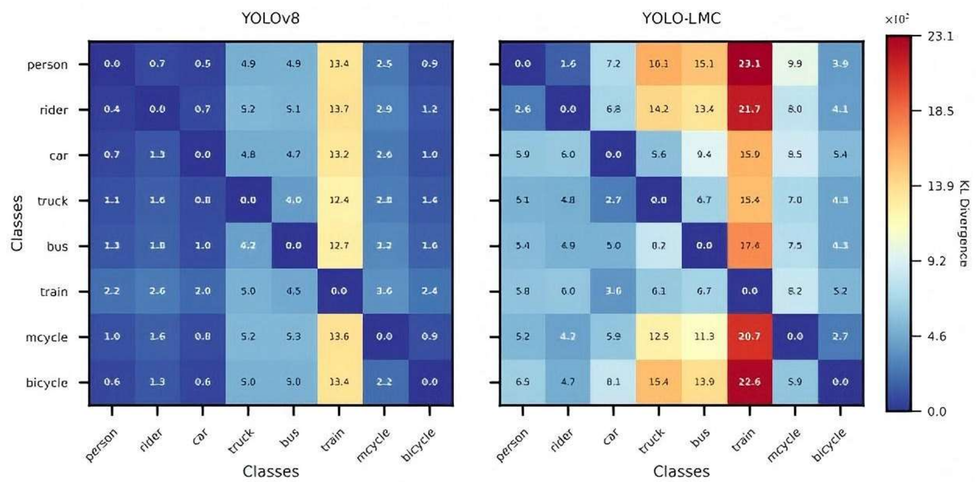

Figure 3. Inter-class KL Divergence on the Cityscapes Dataset

ISSN: 2414-1895 DOI: 10.6919/ICJE.202607_12(7).0010

The proposed YOLO-LMC was used for training, with an input image size of 640x640. Image augmentation was performed using random flipping, translation, and erasing. The stochastic gradient descent optimizer (SGD) [27] was used for optimization, with a weight decay of $5e^{-4}$, momentum of 0.937, batch size of 16, and an initial learning rate of 0.01. The learning rate was decreased to 1e-4 in a cosine manner after a total of 100 epochs. Training was conducted on a single RTX 4090 GPU. To verify the clustering effect of the proposed method in the latent space, the KL Divergence between different classes in the latent space was calculated. As shown in Figure 3, the inter-class KL Divergence on the validation set of the fog-free Cityscapes dataset is presented. It can be seen that after using YOLO-LMC, the KL Divergence between different classes increases, indicating that the distribution differences between different classes in the latent space become larger.

As shown in Figure 4, the inter-class KL Divergence on the validation set of the Foggy-Cityscapes dataset is presented. It can be seen that YOLO-LMC still maintains a large KL Divergence between different classes. In contrast, the original YOLOv8 has a smaller inter-class KL Divergence, indicating that YOLO-LMC can still maintain the differences in class distributions when domain changes occur.

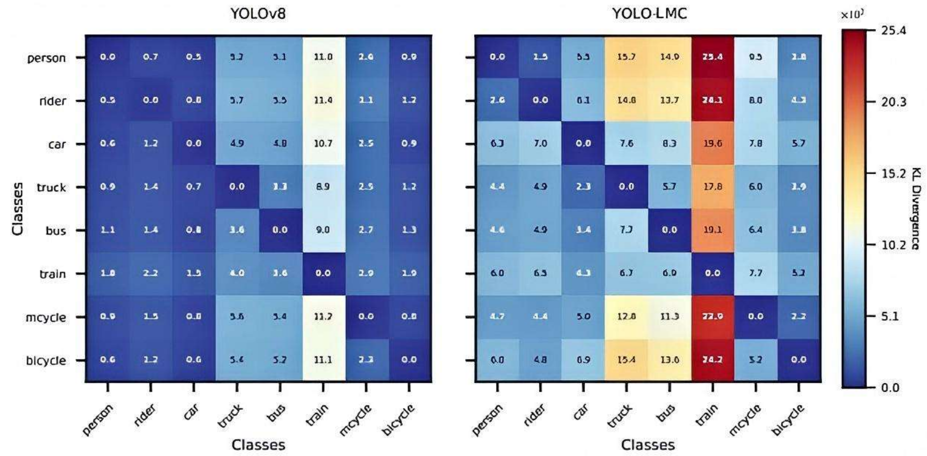

Figure 4. Inter-class KL Divergence on the Foggy-Cityscapes Dataset

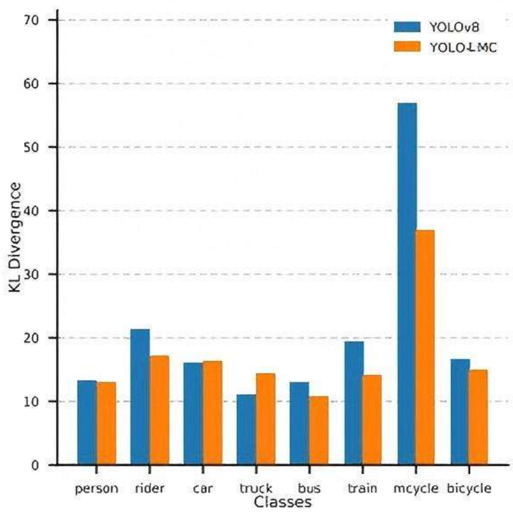

Figure 5. KL Divergence between the Same Classes on the Cityscapes and Foggy-Cityscapes Datasets

ISSN: 2414-1895
DOI: 10.6919/ICJE.202607_12(7).0010

Figure 5 shows the KL Divergence between the same classes on the Cityscapes and Foggy-Cityscapes datasets when approximating the distribution of Foggy-Cityscapes with that of Cityscapes. It can be seen that, except for two classes with slightly larger KL Divergence, the KL Divergence of most classes is much smaller than that of the original YOLOv8. This indicates that YOLO-LMChas a smaller prediction distribution shift between different domains, proving the effectiveness of the proposed method.

To further verify the clustering effect of the proposed method in the latent space, t-SNE[28] was used to reduce the dimensionality of high-dimensional data and visualize it on the Cityscapes and Foggy-Cityscapes datasets. Figure 6 shows the t-SNE clustering visualization without LMC. The points of different classes are mixed together. The left figure in Figure 6 is the visualization on Cityscapes, and the right figure is the visualization on Foggy-Cityscapes. It can be seen that when transitioning from a fog-free scene to a hazy scene, the feature points change significantly. Figure 7 shows the t-SNE clustering visualization with LMC. The points of different classes are effectively separated, and even points that are difficult to separate, such as pedestrians, riders, cars, and trucks, are well separated. The left figure in Figure 7 is the visualization on Cityscapes, and the right figure is the visualization on Foggy-Cityscapes. It can be seen that when transitioning from a fog-free scene to a hazy scene, the feature points change less significantly.This proves the effectiveness of the proposed method.

Figure 6. t-SNE Clustering Visualization without LMC

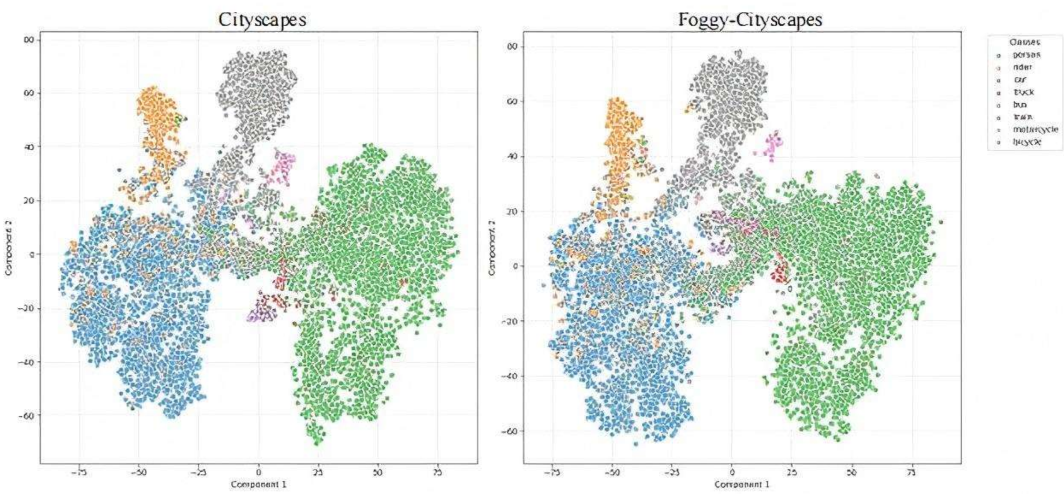

Figure 7. t-SNE Clustering Visualization with LMC

ISSN: 2414-1895 DOI: 10.6919/ICJE.202607_12(7).0010

3.4 Comparison Experiment Results and Analysis

Six advanced cross-domain adaptive object detection methods were selected for comparison. The comparison methods include THNet[16],SW-DA[15], SC-DA[12], IA- YOLO[29], SSA-DA[13], and MS-DAYOLO[14], which are domain-adaptive algorithms.

These algorithms enhance the robustness of the model in cross-domain scenarios through different technical approaches and can be used for a comprehensive comparison of the performance of the algorithm in this paper. The methods without publicly available source code directly referenced data from the official papers, such as THNet, SW-DA, SC-DA, and SSA-DA. The others were experimented on the same RTX 4090 device with the default parameters provided by the official sources.

As shown in Table 2, compared to Faster R-CNN-based methods, YOLO-LMC demon- strated an improvement of 8.8 to 17.5 percentage points in mAP_{50} (details in Table 2). This superior performance can be attributed to the more advanced model structure of the YOLO series and the enhanced anti-interference capability conferred by the latent space metric mechanism, which operates at the level of intrinsic feature similarity.When compared with other YOLO improvement series, such as MS-DAYOLO and IA-YOLO, our method achieved an increase of 9.4 and 8.6 percentage points in mAP_{50} respectively, while main-taining a comparable number of parameters. This underscores the superiority of the metric constraint strategy over mere network structure adjustments.

Table 2. Experiment Results on the Foggy-Cityscapes Dataset.

Method	model	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
SC-DA	F-RCNN	33.5	38	48.5	26.5	39	23.3	28	33.6	33.8
SW-DA	F-RCNN	29.9	42.3	43.5	24.5	36.2	32.6	30	35.3	34.3
THNet	F-RCNN	36.0	44.7	55.1	31.8	48.5	47.9	25.6	40.1	41.2
SSA-DA	F-RCNN	33.9	48.3	47.7	35.7	52.0	44.7	39.6	37.9	42.5
MS-DAYOLO	YOLOv4	44.5	50.4	59.3	28.3	49.1	29.7	29.9	44.2	41.9
IA-YOLO	YOLOv8	50.3	56.7	60.4	22.3	47.2	20.5	38.9	45.6	42.7
YOLO-LMC(ours)	YOLOv8	54.8	60.5	70.4	30	58	50.4	36.7	49.8	51.3

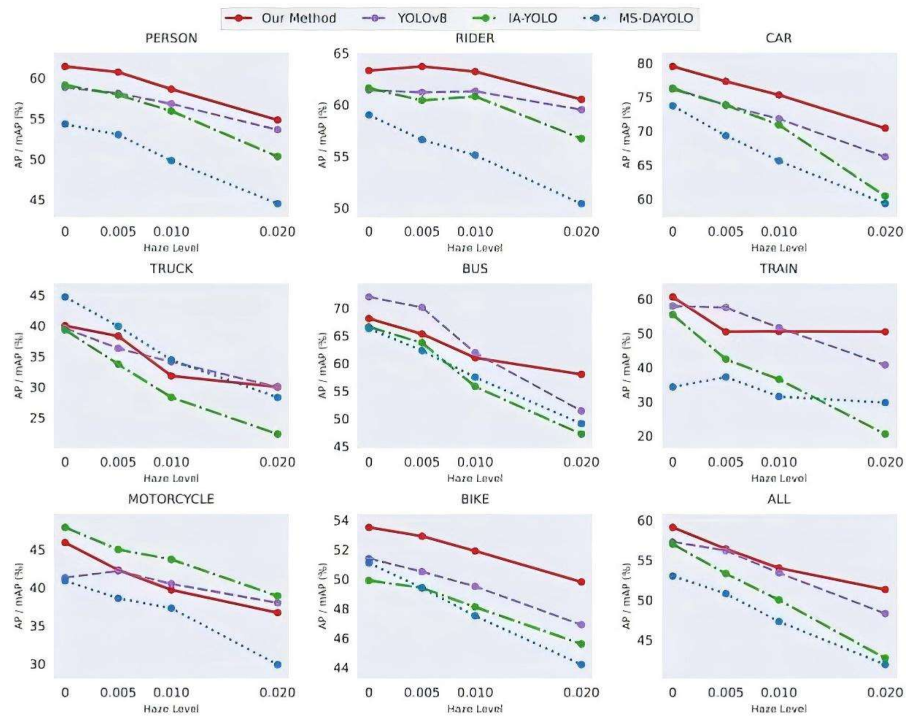

Figure 8. Comparison of AP/mAP under different haze concentrations on Foggy-Cityscapes.

Furthermore, to analyze object detection performance under varying haze concentrations, Figure 8 illustrates the precision changes of several representative methods across different haze densities. The x-axis represents the atmospheric scattering coefficient β from the atmospheric scattering model, and the y-axis denotes the AP for each class or the mAP for all classes. It can be observed that, with the exception of ’truck’, ’bus’, ’train’, and ’motorcycle’, which are classes with relatively few instances in the dataset, YOLO-LMC outperforms other models in terms of AP for other categories and overall mAP. The subop-timal performance on classes with fewer instances is speculated to be due to insufficient sample size, leading to underfitting.

4. Discussion

4.1 Parameter Sensitivity Analysis

To investigate the optimization mechanism of key parameters in the YOLO-LMC model, a systematic parameter sensitivity analysis was conducted for the margin parameter and the number of sub-cluster centers K in the metric loss function. Tables 3 and 4 present the trend of mAP changes on the Foggy-Cityscapes validation set under different parameter values. Experimental results indicate that the model performance reaches its optimal state, forming a clear performance peak, when the margin is set to 1 and K = 12.

Table 3. mAP on Foggy-Cityscapes dataset with different margin values.

margin	**0.5**	**0.75**	**1**	**1.25**	**1.5**
mAP	51.4	51.7	51.7	51.7	51

Table 4. mAP on Foggy-Cityscapes dataset with different K values.

K	**8**	**10**	**12**	**14**	**16**
mAP	50.5	51.3	51.7	50.1	49.3

This parameter sensitivity characteristic can be explained from the theoretical frame-work of metric learning: when the margin is too small, the inter-class repulsive force is insufficient to counteract feature confusion caused by haze, making it difficult for the model to establish clear decision boundaries, leading to suboptimal transfer performance in hazy scenarios. Conversely, when the margin is too large, overly strong inter-class constraints can disrupt the intrinsic structure of the feature space, causing excessive compression of intra-class features and information loss, which is detrimental to generalization in the hazy target domain.

The parameter sensitivity of the K value reflects the delicate balance in modeling intra-class diversity. Experiments reveal that this parameter significantly impacts transfer performance in hazy scenes. If K is set too large, it can lead to redundant cluster centers, increasing the difficulty of model optimization. Conversely, if K is too small, the model may fail to capture the multi-scale morphological variations within the same class. When the number of sub-clusters K=12, the model effectively captures multi-scale intra-class variations while avoiding optimization difficulties due to redundant cluster centers, thus achieving optimal performance.

4.2 Ablation Study and Analysis

A systematic ablation study (Table 5) was conducted to validate the contribution of each innovative module. In this study, "Euc" denotes using Euclidean distance for the metric approach, "Cos" denotes using cosine distance, and "MTAL" and "LMC" are the Mask Task Alignment Learning and Latent Space Metric Constraint algorithms proposed in this paper, respectively. The baseline model A, without any improvements, achieved only 48.3% mAP in hazy scenarios. The application of the cosine distance and LMC metric module (Model B) yielded a significant improvement of 2.6%, confirming the importance of latent space structural constraints. Model D, incorporating MTAL for positive sample screening, further contributed an additional 0.4%, indicating that the dynamic threshold mechanism effectively filters low-quality samples and reduces noise introduction. When MTAL and LMC are jointly optimized, they produce a significant synergistic effect, with mAP_{50} reaching 51.3%, a 3% improvement over the baseline. Model C replaced the cosine distance metric criterion with Euclidean distance; although it showed an improvement over the baseline model, its performance was far inferior to that using cosine distance, demonstrating the correctness of our chosen metric method.

Table 5. Ablation study of YOLO-LMC.

Method	Euc	Cos	MTAL	LMC	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
A					53.6	59.5	66.2	30	51.4	40.7	38	46.9	48.3
B		✓		✓	53.5	59.6	67.8	35.1	54.7	45.1	44	47.7	50.9
C	✓		✓	✓	52.8	57.8	66.2	33.3	55.6	44.9	36.2	46.8	49.2
D		✓	✓	✓	54.8	60.5	70.4	30	58	50.4	36.7	49.8	51.3

5. Conclusion

This paper introduces YOLO-LMC, a novel object detection framework specifically designed to enhance performance in hazy environments by integrating latent-space metric learning. Recognizing that haze degrades image quality and consequently impairs the feature extraction capabilities of neural networks, we proposed a solution that focuses on optimizing the discriminative power of features directly within the latent space.

Our key contributions include:

(1) A Masked Task Alignment Learning (MTAL) strategy for latent space optimization, which intelligently filters high-confidence positive samples, thereby reducing noise and improving the quality of feature selection for metric learning.

(2) A novel category-center-based metric learning architecture utilizing cosine similarity. This design compels intra-class features to form compact clusters around their respective class centers and maximizes the separation between different class centers in the latent space.

Extensive experimental validation was conducted on the Foggy-Cityscapes dataset. The results demonstrate that YOLO-LMC significantly outperforms existing object detection enhancement methods in hazy scenes, achieving an impressive improvement of at least 8.6% in mAP. Ablation studies further confirm the effectiveness of each proposed module, particularly highlighting the synergistic benefits of combining MTAL with the latent-space metric constraints.

While the proposed YOLO-LMC achieved state-of-the-art performance, we acknowledge that its performance on classes with very few instances (e.g., ‘truck’, ‘bus’, ‘train’, ‘motorcycle’) could be further improved. This is likely due to the inherent data imbalance in the dataset, leading to potential underfitting for these sparse categories. Future work will explore strategies to address this, such as few-shot learning techniques or advanced data augmentation specific to underrepresented classes in hazy conditions.

Author Contributions

Conceptualization, G.L. and S.L.; methodology, G.L. and S.L.; software, C.J.; validation, Y.M., X.W. and C.J.; formal analysis, C.J.; investigation, Z.Y. and Y.L.; resources, G.L. and S.L.; data curation, C.J.; writing-original draft preparation, C.J.; writing-review and editing, H.L., W.Z. and C.J.; visualization, C.J.; supervision, W.Z.; project administration, H.L.; funding acquisition, G.L. and S.L All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement

Not applicable.

Informed Consent Statemen

Not applicable.

Data Availability Statement

The dataset and source code that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

Funding: This research was supported by Zhejiang Zheneng Jinhua Gas Turbine Power Generation Co., Ltd. (Contract No. JHGT202412048-SG), for the project titled "Research on 9E Gas Turbine Combustion Pressure Pulsation Monitoring and Combustion Adjustment".

References

[1] Nazir, A., & Wani, M. A. (2023). You only look once-object detection models: A review. In *Proceedings of the 10th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1088–1095). IEEE.

[2] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 1137–1149.

[3] Yang, S., Lu, H., & Li, J. (2022). Multifeature fusion-based object detection for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 24, 1126–1133.

[4] Feng, D., Harakeh, A., Waslander, S. L., & Dietmayer, K. (2021). A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23, 9961–9980.

[5] Abba, S., Bizi, A. M., Lee, J. A., Bakouri, S., & Crespo, M. L. (2024). Real-time object detection, tracking, and monitoring framework for security surveillance systems. *Heliyon*, 10.

[6] Gupta, H., Kotlyar, O., Andreasson, H., & Lilienthal, A. J. (2024). Robust object detection in challenging weather conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 7523–7532).

[7] Li, C., Zhou, H., Liu, Y., Yang, C., Xie, Y., Li, Z., & Zhu, L. (2023). Detection-friendly dehazing: Object detection in real-world hazy scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 8284–8295.

[8] Mullappilly, S. S., Gehlot, A. S., Anwer, R. M., Khan, F. S., & Cholakkal, H. (2024). Semi-supervised open-world object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, pp. 4305–4314).

[9] Husain, A. A., Maity, T., & Yadav, R. K. (2020). Vehicle detection in intelligent transport system under a hazy environment: A survey. *IET Image Processing*, 14, 1–10.

[10] Sharma, T., Nalla, B. T., Verma, N. K., & Vasikarla, S. (2022). FR-HDNet: Faster RCNN based haze detection network for image dehazing. In *Proceedings of the 2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)* (pp. 1–8). IEEE.

[11] Cai, B., Xu, X., Jia, K., Qing, C., & Tao, D. (2016). Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25, 5187–5198.

[12] Zhu, X., Pang, J., Yang, C., Shi, J., & Lin, D. (2019). Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 687–696).

[13] Zhao, Z., Guo, Y., & Ye, J. (2020). Bi-dimensional feature alignment for cross-domain object detection. In *Proceedings of the European Conference on Computer Vision* (pp. 671–686). Springer.

[14] Hnewa, M., & Radha, H. (2021). Multiscale domain adaptive YOLO for cross-domain object detection. In *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)* (pp. 3323–3327). IEEE.

[15] Saito, K., Ushiku, Y., Harada, T., & Saenko, K. (2019). Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6956–6965).

[16] Song, W., Ren, S., Tan, W., & Wang, X. (2024). THNet: Transferability-aware hierarchical network for robust cross-domain object detection. *IEEE Access*.

[17] Varghese, R., & Sambath, M. (2024). YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In *Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)* (pp. 1–6). IEEE.

[18] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2117–2125).

[19] van Doorenmalen, J., & Menkovski, V. (2020). Evaluation of CNN performance in semantically relevant latent spaces. In *Proceedings of the International Symposium on Intelligent Data Analysis* (pp. 145–157). Springer.

[20] Deng, L., Li, H., Liu, H., & Gu, J. (2022). A lightweight YOLOv3 algorithm used for safety helmet detection. *Scientific Reports*, 12, 10981.

[21] Feng, C., Zhong, Y., Gao, Y., Scott, M. R., & Huang, W. (2021). TOOD: Task-aligned one-stage object detection. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3490–3499). IEEE Computer Society.

[22] Wang, G., Li, S., Zhu, X., Wang, Y., Huang, J., Zhong, Y., & Wu, Z. (2025). REW-YOLO: A lightweight box detection method for logistics. *Modelling*, 6.

[23] Du, S., Zhang, B., Zhang, P., & Xiang, P. (2021). An improved bounding box regression loss function based on CIOU loss for multi-scale object detection. In *Proceedings of the 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)* (pp. 92–98). IEEE.

[24] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3213–3223).

[25] Sakaridis, C., Dai, D., & Van Gool, L. (2018). Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126, 973–992.

[26] Joyce, J. M. (2011). Kullback-leibler divergence. In S. Kotz, C. B. Read, & N. L. Johnson (Eds.), *International encyclopedia of statistical science* (pp. 720–722). Springer.

[27] Cui, X., Zhang, W., Tüske, Z., & Picheny, M. (2018). Evolutionary stochastic gradient descent for optimization of deep neural networks. *Advances in Neural Information Processing Systems*, 31.

[28] Linderman, G. C., & Steinerberger, S. (2019). Clustering with t-SNE, provably. *SIAM Journal on Mathematics of Data Science*, 1, 313–332.

[29] Liu, W., Ren, G., Yu, R., Guo, S., Zhu, J., & Zhang, L. (2022). Image-adaptive YOLO for object detection in adverse weather conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 1792–1800).