Pattern Recognition 179 (2026) 113516

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Image signal process with dynamic class-rebalanced and IoU-threshold for unsupervised domain adaptive dark object detection

Yin Zhang [a,b,1], Yongqiang Zhang [c,1,*], Zian Zhang [a], Man Zhang [a], Rui Tian [a], Mingli Ding [a], Bogdan Raducanu [b], Dan Liu [a,*]

[a] School of Instrumentation Science and Engineering, Harbin Institute of Technology, Harbin, 15001, China
[b] Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, 08193, Spain
[c] College of Computer Science, Inner Mongolia University, Hohhot, 010021, China

ARTICLE INFO

Keywords:
Dark object detection
Image signal process
Unsupervised domain adaptive

ABSTRACT

Object detection in dark conditions has always been a great challenge due to the complex formation process of low-light images. Currently, the mainstream methods usually adopt domain adaptation with Teacher-Student architecture to solve the dark object detection problem, and they imitate the dark conditions by using non-learnable data augmentation strategies on the annotated source daytime images. Note that these methods neglect to model the intrinsic imaging process, *i.e.* image signal processing (ISP), which is important for camera sensors to generate low-light images. Furthermore, the class imbalance in the dataset often leads to incorrectly detecting a minority class as a majority class, which exerts a huge impact on dark object detection as well. To solve the above problems, in this paper, we propose a novel method named ISP Dynamic Teacher for dark object detection by exploring Teacher-Student architecture from a new perspective (*i.e.* self-supervised learning based ISP degradation, dynamic class-rebalanced and IoU-threshold adjustment strategy). Specifically, we first crop all the instances of each image in a batch size according to their ground truth and save them into a Mix-Instances-Bank. Then, we propose a Class-Rebalanced Module (CRM) by creating a confusion matrix and use a weight sampling algorithm to identify the mixed instance from the Mix-Instance-Bank that best matches the base instance in the original source image. Subsequently, by mixing them together with the help of Mixup augmentation, we can address the problem of class imbalance in object detection. Moreover, we design a Day-to-night Transformation Module that is consistent with the ISP pipeline of the camera sensors (ISP-DTM) to make the augmented images look more in line with the natural low-light images captured by cameras, and the ISP-related parameters are learned in a self-supervised manner. Then, to avoid the conflict between the ISP degradation and detection tasks in a shared encoder, we propose a disentanglement regularization (DR) that minimizes the absolute value of cosine similarity to disentangle two tasks and push two gradient vectors as orthogonal as possible. Building upon this, we further devise a Dynamic IoU-threshold Adjustment (DIA) strategy to find the True-Positive (TP) pseudo-labels that are erroneously filtered out by a high static threshold. Extensive experiments show that we achieve state-of-the-art performance in dark object detection on the BDD100k dataset (49.6% in AP) and the SHIFT dataset (52.8% in AP) respectively. The code can be found at https://github.com/zhangyin1996/ISP-Teacher.

1. Introduction

Object detection has achieved remarkable success and is widely used in various fields such as autonomous driving [1,2]. However, these models trained on high-quality daytime images often perform poorly on low-light images, because these images taken in dark conditions suffer from various types of light and undesirable noise. Furthermore, it is also challenging to annotate low-light images. As a result, it is infeasible to acquire high-quality annotation information for low-light images in the same way as that for daytime images. Currently, robust and accurate object detection under dark conditions has become increasingly essential for safety-critical fields, yet it remains highly challenging due to low visibility and sensor noise.

A simple way to solve this problem is to perform dark enhancement on low-light images firstly, and then send them to an off-the-shelf detector for object classification and regression. Unfortunately, enhanced

* Corresponding authors.
E-mail addresses: zhangyongqiang@imu.edu.cn (Y. Zhang), liudan@hit.edu.cn (D. Liu).
[1] The two authors contribute equally to this work.

https://doi.org/10.1016/j.patcog.2026.113516
Received 6 April 2025; Received in revised form 9 March 2026; Accepted 15 March 2026
Available online 17 March 2026

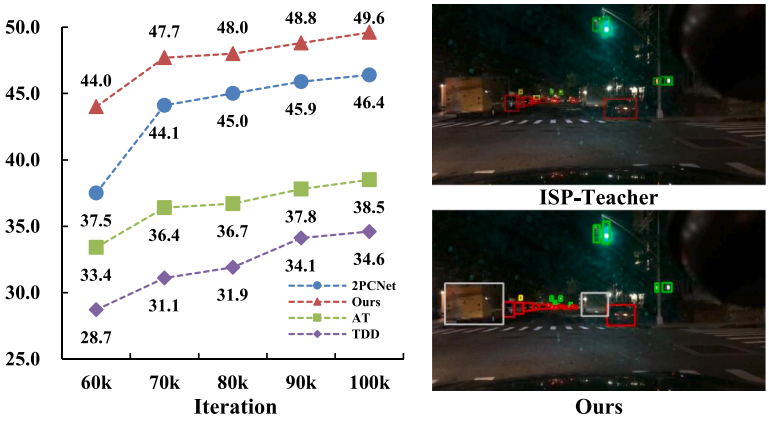

Fig. 1. Results of some teacher-student architecture UDA methods on BDD100k. We found that AT [3] and TDD [4] get worse results on day-to-night conditions, which even lower than the baseline detector Faster-RCNN [5] (41.1% AP). Our proposed method outperforms other counterparts by a large margin and always higher other methods in any iteration. Moreover, our method can also detect the minority class in the dataset, *e.g.* 'Truck' (white boxes) and do not miss detecting small objects, *e.g.* 'Traffic Light' (green boxes) in the distance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

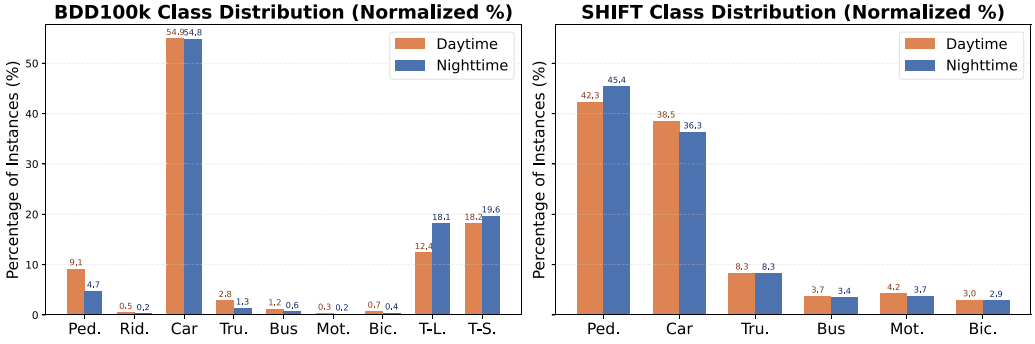

Fig. 2. Class instance distribution of BDD100k and SHIFT dataset in day-to-night domain adaption (Normalized %). In our experiments, there are 9 classes in BDD100k, and the full classes name from left to right are Pedestrian, Rider, Car, Truck, Bus, Motorcycle, Bicycle, Traffic Light and Traffic Sign. And there are 6 classes in SHIFT, which are included in BDD100k.

images are visually comfortable for humans but do not benefit from the high-level task of machine vision [6,7]. Recently, the teacher-student architecture has attracted a lot of attention in semi-supervised object detection [3,8,9] and has also achieved excellent results in the field of domain adaptation object detection [4,10,11]. However, as shown in Fig. 1, we found that the best teacher-student UDA methods like AT [10], TDD [4] achieve good results on regular domain adaptive datasets (*e.g.* Cityscapes to Foggy Cityscapes) but get poor results under day-to-night conditions. They are even lower than the baseline detector (*i.e.* Faster-RCNN [5]) that is trained on daytime images and applied directly to nighttime images (41.1% AP). Therefore, although UDA has been proposed to address this problem, existing UDA methods still face several limitations and challenges under dark conditions:

Datasets commonly used in dark object detection tasks (*e.g.* BDD100k [12] and SHIFT [13]) have a serious class imbalance problem. As shown in Fig. 2, these two datasets have completely different class distributions. For BDD100k dataset, 'Car' is a majority class and 'Pedestrian' is a minority class, but both 'Car' and 'Pedestrian' classes are majority classes for the SHIFT dataset. Moreover, the gap between the majority classes and the minority classes in the two datasets is very large, which leads to poor performance in day-to-night domain adaption detection.

ii) Most of these teacher-student based methods solve the domain bias problem by optimizing the framework [4] or selecting useful ground-truth labels [8]. They usually imitate the dark conditions by using traditional non-learnable data augmentation strategies on the available source daytime images. However, these methods neglect to model the intrinsic imaging process (ISP) of camera sensors and may distort the physical properties of the scene, thereby introducing domain bias to the detector. In contrast, our learnable ISP module models the real camera ISP pipeline (*e.g.* tone mapping, gamma correction, noise modeling, and white balance) in a self-supervised manner. This enables the model to capture intrinsic visual information that remains consistent across domains and improves adaptation in the UDA setting.

iii) Most of the teacher-student UDA methods utilize a static threshold to generate pseudo-labels. However, these obtained pseudo-labels might be highly inaccurate [3], *i.e.* while low confidence pseudo-labels are filtered out, the True-Positive pseudo-labels are filtered out simultaneously as well. Furthermore, small objects or minority classes are already hard to see in dark conditions, and a higher static threshold would also cause them to be further lost [14], which in turn makes it more difficult for the model to learn representations of these objects during training. As shown in Fig. 1, ISP-Teacher [15] misses the detection of small objects, *i.e.* 'Traffic Light' (green boxes) in the distance.

In this article, the goal of our work is to construct a robust teacher-student framework under the day-to-night domain shift by jointly leveraging a learnable ISP-based degradation module and dynamic class-rebalanced and IoU-threshold strategies. Note that this paper extends our previous work ISP-Teacher [15] published in AAAI 2024. In ISP-Teacher, we explore teacher-student architecture from a novel perspective of self-supervised learning based ISP degradation for dark object detection. More specifically, we study how to use self-supervised learning to capture intrinsic visual information that is not affected by lighting changes, which can address the domain bias of the student network.

Compared with ISP-Teacher, we make two extensions in this paper: (1) a class-rebalanced module (CRM) is proposed to address the problem of class imbalance in dark object detection; and (2) a dynamic IoU-threshold adjustment (DIA) is designed to improve the quality of pseudo-labels. The improved method is named ISP Dynamic Teacher, and to sum up, the contributions of this paper are as follows:

- We first propose a Class-rebalanced Module (CRM) by creating a confusion matrix and utilize a weight sampling algorithm to find the mixed instance that best matches the base instance in the original source image. And then mixing them together uses Mixup augmentation, which could solve the problem of class imbalance in dark object detection.
- We design a day-to-night transformation module (ISP-DTM) inspired by the image signal processing pipeline of camera sensors to generate dark images from daytime images, and the obtained dark images are compatible with the natural low-light images captured by the camera, which can address the domain bias of the student network.
- Moreover, a disentanglement regularization is imposed by minimizing the gradients of cosine similarity of two different tasks (*i.e.* self-supervised learning based ISP degradation and object detection) while maximizing cosine similarity of the same tasks. This simple implement could decouple these two tasks in a shared encoder.
- Finally, we introduce a Dynamic IoU-threshold Adjustment (DIA) to adjust the threshold dynamically for finding True-Positive pseudo-labels by calculating the IoU of teacher and student bounding-box predictions, which further improving the quality of pseudo-labels.
- Extensive experiments conducted on BDD100k and SHIFT datasets show the effectiveness of our proposed method. In particular, ISP Dynamic Teacher achieves the new SOTA performance on BDD100k (49.6% in AP) and SHIFT datasets (52.8% in AP), respectively.

2. Related work

Object Detection in Dark Conditions. To tackle the problem of object detection in low-light conditions, a direct way is to use low-light enhancement methods [16] to process the dark images and then send the de-dimming images to the mainstream object detection methods [5,17] for inference. However, the detection performance of these methods is unsatisfactory in some natural dark images. As a result, there are some end-to-end methods that train the low-light enhancement and object detection tasks jointly. For example, IA-YOLO [18] designs a filter module with a learnable parameter trained jointly with YOLOv3 in an end-to-end fashion to balance the tasks of image enhancement and object detection. MAET [7] introduces a multitask auto-encoding transformation model to decode low-light degradation transformation considering noise and the ISP pipeline in cameras. IAFE [19] addresses low-light object detection by jointly optimizing illumination-adaptive enhancement and detection with dynamic feature interaction.

Recently, 2PCNet [11] makes its first attempt at Day-to-Night unsupervised domain adaptive object detection. They design an image augmentation pipeline called NightAug to solve the problems that come from low-light regions and other night-related attributes in images. Furthermore, Cooperative Student (Cos) [20] proposes a global-local transformation that transfers the prior knowledge from source to target domain and addresses the inconsistencies across various regions of the same object captured at nighttime. Debiased-Teacher [21] proposes a debiasing framework, which alleviates distribution, training, and confirmation biases via domain transforming, cross-domain representation compensation, and pseudo-label calibration. However, these methods propose a non-learnable data augmentation while our method is self-supervised and we consider the principle of the intrinsic imaging process of camera sensors.

Class-rebalanced in High-level Tasks. Class imbalance has attracted the attention of many scholars, and some solutions have been proposed. For example, CutMix [22] randomly crops some patches from one image and pastes them onto another image to make the network have better generalization and object localization capabilities. In addition, OpenReMix [23] is designed for the open-set domain adaptation task, and it resizes a random class from the source image and mixes it with the target image, at the same time, it also cuts the parts from a target image and pastes them into a source image to learn size-invariant features.

However, the above methods either randomly crop images or paste the image at a random location without considering the relationship between the classes in the dataset. Our proposed class-rebalanced module (CRM) takes into account the relationship between classes and finds the mixed instance that best matches the base instance in the original source image based on a confusion matrix and a weight sampling algorithm.

Disentanglement Regularization. Multi-task networks usually contain an encoder and several decoders for specific tasks. However, these approaches face an optimization problem which sometimes leads to worse performance than training each task independently. At present, scholars generally believe the main reason for this phenomenon is gradient conflict, and some methods have been proposed to solve this problem. For instance, [24] alters the gradients by projecting the gradient of one task onto the normal plane of the gradient of the other task when the values of cosine similarity are negative[25]. finds nearly orthogonal gradients would not interfere with each other's tasks, and proposes that regularizing the angle between gradients to solve the negative transfer problem.

The above methods are focused on classification and regression tasks, and our work is inspired by recent research [7] that minimizes the absolute value of cosine similarity to disentangle the object detection and degrade transformation tasks. Different from [7], we design a simple disentanglement regularization to decouple our self-supervised learning based ISP degradation and detection tasks in the teacher-student architecture by minimizing the cosine similarity of different tasks while maximizing the cosine similarity of the same tasks.

Dynamic IoU-threshold Adjustment Strategy. Currently, there are some methods that propose to revise the score threshold of pseudo-labels dynamically during training. For example, Consist-Teacher [3] proposes a Gaussian Mixture Model (GMM) to flexibly determine the threshold for each class, which addresses the oscillating pseudo-targets that undermine the training in the semi-supervised object detection task. [14] introduces a novel low confidence pseudo label distillation (LPLD) loss to mine *hard-positive* samples (*i.e.* minor classes and small objects) that are ignored in low confidence pseudo-labels. By this way, the number of pseudo-labels is complemented and the network gains a deeper understanding of the target domain.

In addition, Cos [20] combines proxy-based target consistency (PTC) and adaptive IoU informed thresholding (AIT) to dynamically update the threshold and capture useful pseudo-labels as learning advances. However, this method requires building a proxy network in advance and then calculating the IoU between this proxy network and the teacher model, which makes the model redundant and affects the computational efficiency of the model. In this paper, we improve the Cos [20] method by calculating IoU of bounding-box predictions between teacher and student models directly to update the threshold without the need for a proxy network.

3. Proposed method

In this section, we describe the details of the proposed ISP dynamic teacher for unsupervised domain adaptive dark object detection. First, we present an overview of the proposed novel teacher-student architecture, as shown in Fig. 3. Then, a class-rebalanced module (CRM) based confusion matrix and weight sampling algorithm are proposed to solve the class imbalance problem. Self-supervised learning based ISP degradation that contains a day-to-night transformation module (ISP-DTM) and a self-supervised learning strategy is proposed from a new perspective to explore the detection task in challenging low-light

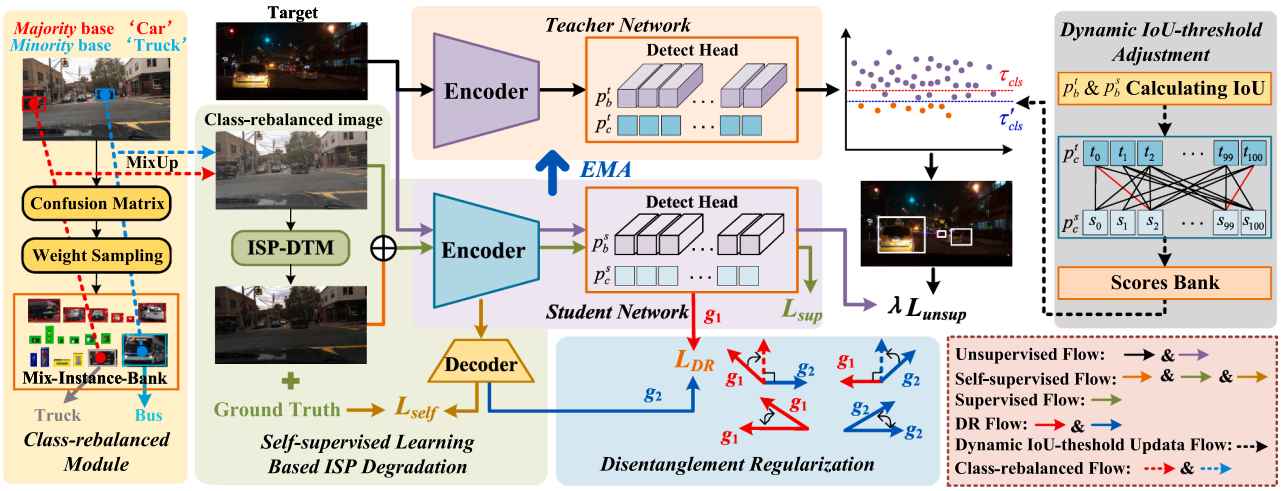

Fig. 3. The pipeline of our ISP Dynamic Teacher. **(A)** The yellow area (left side) is the Class-rebalanced Module (CRM), and the majority base classes and minority base classes from source images are matched to the most relevant mixed class in the Mix-Instance-Bank through confusion matrix and weight sampling algorithm. Then, we resize the mixed instance to the same size as the base instance and use the Mixup algorithm to obtain the class-rebanlanced image. **(B)** The green area is the proposed self-supervised learning based ISP degradation. We first use a day-to-night transformation-fusion module (ISP-DTM) to generate dark images from daytime class-rebalanced images. Then, we calculate L_{self} between ground truths and degradation predictions of an Encoder-Decoder output for self-supervised learning. **(C)** The blue area (lower part) is the illustration of disentanglement regularization (DR), which is used to address the problem of conflicting gradients. The red arrow g_1 and blue arrow g_2 denote gradient vectors of the task of object detection and self-supervised learning based ISP degradation respectively. **(D)** The gray area (right side) is the Dynamic IoU-threshold Adjustment (DIA). We calculate IoU between the prediction bounding boxes p_b^t of the teacher model and prediction bounding boxes p_b^s of the student model. Only the IoU values that are greater than an IoU threshold, and the teacher and student models predict the same class (*e.g.* red line t_0 with s_2 and t_{100} with s_{99}) will be saved in the Scores Bank. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conditions. Moreover, disentanglement regularization (DR) is designed to overcome the problem of conflicting gradients in our pipeline. Finally, a dynamic IoU threshold adjustment (DIA) strategy is proposed to mine high-quality pseudo-labels for the teacher-student architecture.

3.1. Overview of ISP dynamic teacher

Let $D_{day} = \{X_l, Y_l\}$ denotes the daytime dataset, which contains X_l daytime images with Y_l labels in the source domain. $D_{night} = \{X_u\}$ denotes the nighttime dataset, and it only contains X_u nighttime images without labels in the target domain. The subscripts l and u indicate the labeled and unlabeled data, respectively. As shown in Fig. 3, ISP Dynamic Teacher consists of a student network and a teacher network. Similar to prior works [9,11], both student and teacher are the Faster-RCNN [5] structure, and the detection loss L_{det} is as:

$$L_{det} = L_{sup} + \lambda L_{unsup} \tag{1}$$

where L_{sup} and L_{unsup} denote the supervised learning loss and unsupervised learning loss, respectively. λ is a hyper-parameter.

The training process of our method is that the teacher network generates pseudo-labels Y_p to train the student while the student updates the teacher network with exponential moving average (EMA). First, the student network is burned up on daytime images (*i.e.* source domain) under a supervised manner, and the supervised loss is formulated as:

$$L_{sup} = \frac{1}{N_l} \sum_{i=1}^{N_l} \left[L_{cls}(X_l^i, Y_l^i) + L_{reg}(X_l^i, Y_l^i) \right] \tag{2}$$

where L_{cls} is the classification loss of RPN and ROI head in Faster-RCNN, and L_{reg} is the Smooth L_1 loss for bounding box regression. N_l denotes the number of labeled daytime images.

The role of the teacher network is updated through an EMA of the student model's parameters after each iteration, ensuring stable pseudo-label generation under domain shift. Specifically, the teacher network only takes nighttime images (*i.e.* target domain) as the input, and it is used to produce pseudo-labels for the student with an unsupervised loss:

$$L_{unsup} = \frac{1}{N_u} \sum_{i=1}^{N_u} L_{cls}(X_u^i, Y_p^i) \tag{3}$$

where X_u, Y_p denote unlabeled nighttime images and pseudo-labels, respectively. N_u denote the number of unlabeled nighttime images. Note that the unsupervised loss is only applied in the classification while not used in the bounding-box regression.

Furthermore, there are four components in the proposed ISP Dynamic Teacher. The first component is the class-rebalanced module (CRM, yellow area on the left of Fig. 3), which uses the confusion matrix and the weight sampling algorithm to find the mixed instance that best matches the base instance. The second component is the self-supervised learning based ISP degradation (green area next to the CRM), which is used to capture the intrinsic visual information that is not affected by lighting changes. The third component is the disentanglement regularization (DR, blue area at the bottom of Fig. 3), which disentangles dark object detection and self-supervised learning based ISP degradation to mitigate the impact between each other. The last component is the dynamic IoU-threshold adjustment (DIA, gray area on the right of Fig. 3) strategy that dynamically adjusts the threshold to find true-positive ones among the pseudo-labels. In the next subsection, we will illustrate our proposed class-rebalanced module (CRM), self-supervised learning based ISP degradation, disentanglement regularization (DR) and dynamic IoU-threshold adjustment (DIA) strategy in detail.

3.2. Class-rebalanced module (CRM)

As mentioned above, the class imbalance in the datasets has a huge impact on dark object detection. From some failed detection results, we can find that objects are often detected as similar objects, *e.g.* 'Truck' class in BDD100k dataset is often incorrectly detected as 'Car' class but not as 'Pedestrian' class. The reason is that the 'Truck' class has a higher similarity with 'Car' class, but a lower similarity with the 'Pedestrian' class. Inspired by these results, in this paper, we design a Class-rebalanced Module (CRM) to address the impact of class imbalance on

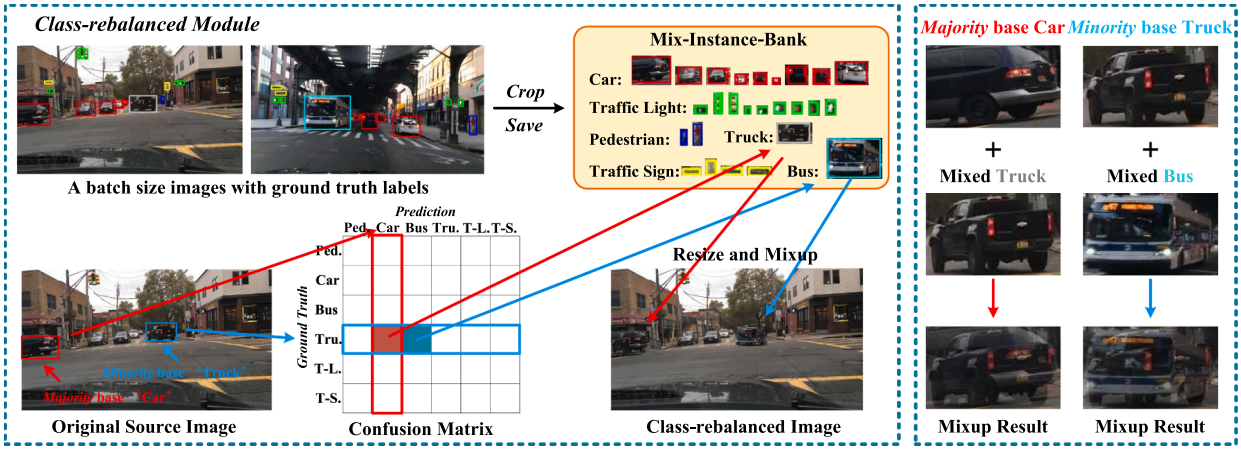

Fig. 4. The structure of Class-rebanlanced Module. For a batch of input images, we use ground truth labels to crop all the instances and save these instances in Mix-Instance-Bank according to class (the orange area in the upper right corner). Then, we utilize Eq. (4) to determine whether each instance in the original source image belongs to the *Majority* base class or the *Minority* base class. Assuming 'Car' is the majority base class and 'Truck' is the minority base class. For the majority base class (or instance), we use the columns of the confusion matrix to select a mixed class (or instance) that is often incorrectly detected as the majority class (or instance), *i.e.* class 'Truck'. Finally, we select a mixed instance 'Truck' from the Mix-Instance-Bank and resize it to the same size as the majority base instance 'Car', and then use the Mixup algorithm to solve the class imbalance problem (the process is shown by the red arrow in the figure). Moreover, the only difference for the minority base class (instance) 'Truck' is that we use the rows of the confusion matrix to select a mixed class (or instance) 'Bus' and the whole process is shown by the blue arrow. The right side of Fig. 4 is the visualization results. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

dark object detection. Different from the previous methods [22,23,26] that increase the number of classes randomly to solve the class imbalance problem, CRM increases the number of some specific classes based on the confusion matrix and weight sampling algorithm.

The preliminary preparation is similar to previous methods [27,28], as shown in the top of Fig. 4, we crop all the instances of each image in a batch size according to their ground truth bounding boxes and save them into a Mix-Instances-Bank. The goal of our proposed CRM is to select a **mixed** instance I_{mix} (named mixed class) from the Mix-Instances-Bank and then mix it with a **base** instance I_{base} (named base class) in the original source image. Therefore, how to choose I_{base} and I_{mix} is the core problem of our Class-rebalanced Module.

Specifically, i) to select the base instance, firstly, we create a confusion matrix $M \in R^{N_c \times N_c}$, where N_c is the number of classes in the dataset, and initialize it with zero. Notes that the confusion matrix is continuously updated as the model forwards under the supervised loss L_{sup}. Each column of the confusion matrix is the prediction that comes from the RoI-Head of the Faster-RCNN detector for each class, and each row is the ground truth of the class. Secondly, we calculate the average value on the diagonal of the confusion matrix $M(c_i, c_i)$ as in Eq. (4), *i.e.* the correct predictions probability of RoI-Head in the Faster-RCNN detector:

$$M_{avg} = \frac{1}{N_c} \sum_{i=0}^{N_c} M(c_i, c_i) \tag{4}$$

where M_{avg} is the average value, c_i and N_c denote the i-th class and the total number of classes in the dataset, respectively. We emphasize that for each ground truth of the class c_i and prediction $\hat{c}_i$, the confusion matrix $M(c_i, \hat{c}_i)$ is updated using a standard EMA during training iterations to maintain stability, and the corresponding matrix is updated as: $M(c_i, \hat{c}_i) \leftarrow M(c_i, \hat{c}_i) + 1$. Following [28], the base instance usually is divided into two categories: (1) $M(c_i, c_i)$ that is above M_{avg} as the **majority** base instance. (2) Similarly, $M(c_i, c_i)$ below the M_{avg} as the **minority** base instance. For example, from Fig. 4, we assume that 'Car' is the majority base class and 'Truck' is the minority base class.

ii) Assume that we have selected a base instance I_{base} (a majority base instance or minority base instance), and its corresponding class is c_i. The next step is to use the confusion matrix M to choose a mixed instance I_{mix} that best matches this base instance I_{base}. One theory is that the columns of the confusion matrix represent the probability of a class being predicted as the current ground truth class, and the rows of the confusion matrix represent the probability of the current ground truth class being represented as other classes.

Therefore, the selection process of mixed instance I_{mix} is as follows: (1) if the base instance I_{base} is a majority base instance, we use the columns of the confusion matrix, *i.e.* $V_{majority} = M(:, c_i) \in R^{1 \times N_c}$, to select a class that is often incorrectly detected as the majority class, and the probability of sampling class c_j is: $P(c_j \mid c_i) = M(c_j, c_i) / \sum_{k=1}^{N_c} M(k, c_i)$. In this case, as shown in Fig. 4, it represents 'Car' (c_i) column (red rectangle) in the Confusion Matrix. (2) And if the base instance I_{base} is a minority base instance, the only difference is that we use the rows of the confusion matrix, *i.e.* $V_{minority} = M(c_i) \in R^{1 \times N_c}$ to select a mixed class, and the probability of sampling class c_j is: $P(c_j \mid c_i) = M(c_i, c_j) / \sum_{k=1}^{N_c} M(c_i, k)$. In this case, it represents 'Truck' (c_i) row (blue rectangle in Fig. 4) in the Confusion Matrix.

Then we apply a weight sampling algorithm to get the mixed class c_{mixed} from $V_{majority}$ or $V_{minority}$. As shown in Fig. 4, for majority class 'Car', c_{mixed} is 'Truck' (red square area in Confusion Matrix) that best matches majority class 'Car', and for minority class 'Truck', c_{mixed} is 'Bus' (blue square area in Confusion Matrix) that best matches minority class 'Truck'. Finally, the mixed instance I_{mix} is selected by c_{mixed} from Min-Instances-Bank. Typically, there may be many instances of the corresponding class in Min-Instances-Bank. In this paper, we select one instance randomly as a mixed instance I_{mix}. Furthermore, we resized the I_{mix} to be the same size as I_{base} by bounding box for not affecting the subsequent detection task.

After obtaining the base instance I_{base} and the resized mixed instance I_{mix}, we use Mixup augmentation to address the problem of class imbalance in dark object detection:

$$\hat{I} = \rho \cdot I_{base} + (1 - \rho) \cdot I_{mix} \tag{5}$$

where $\hat{I}$ is the class-rebalanced image, ρ is the mix ratio, I_{base} is the base instance in the original source image, and I_{mix} is the selected mixed instance that best matches the base instance in the Mix-Instances-Bank. The whole process of the CRM is summarized in Algorithm 1.

Our CRM provides a more focused and principled approach than traditional random sampling for the following key reasons: First, unlike traditional methods such as random sampling or reweighting, which treat

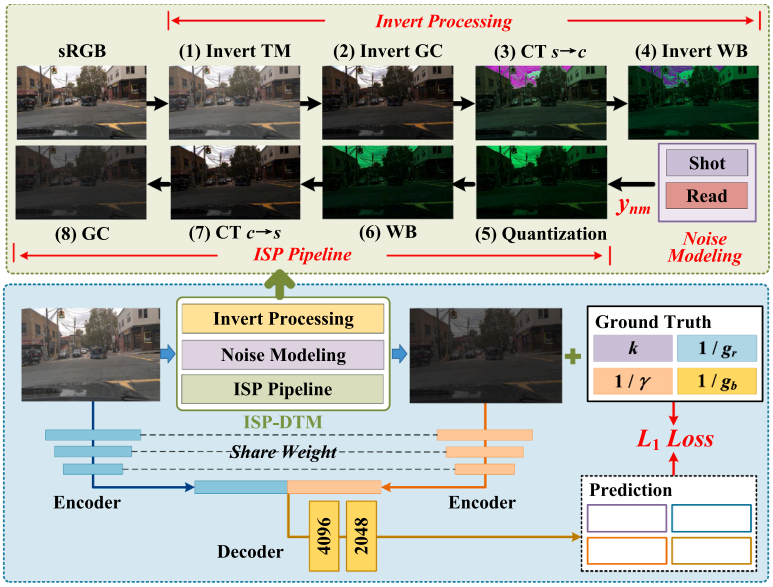

Fig. 5. The structure of self-supervised learning based ISP degradation. Original class-rebalanced daytime images pass through Invert Processing, Noise Modeling, and ISP Pipeline steps to obtain low-light images and four parameters, *i.e.* k, $1/\gamma$, $1/g_r$, $1/g_b$ as ground truths. Then, the weight-shared encoder extracts features from this pair of images and a decoder is used to decode the degradation parameters. Finally, the predictions of the decoder are supervised under the L_1 loss.

Algorithm 1 The Algorithm of Class-rebalanced Module.

Input: A batch size images $Images$ from original input images.
A confusion matrix $M \in R^{N_c \times N_c}$ in which each element is initialized with 0.
Average value M_{avg} calculated by Eq. (4).
while training **do**
  ***Step 1*:** Crop instances and save them in $Bank$.
  for img in batch size images $Images$ **do**
    for gt_box b_i, gt_cls c_i in img **do**
      i) Crop $instance$ from img by using gt_box b_i; ii) Save them in $Bank$: $Bank[c_i][instance]$
    end for
  end for
  ***Step 2*:** Base instance I_{base} and mixed instance I_{mix} selection.
  for img in batch size images $Images$ **do**
    for gt_box b_i, gt_cls c_i in img **do**
      if $M(c_i, c_i) > M_{avg}$ **then**
        i) $V_{majority} = M(:, c_i) \in R^{1 \times N_c}$. The current c_i is the majority base class and the responding instance I_{base} is the majority base instance
        ii) Weight sampling algorithm: c_{mixed} = torch.multinomial($V_{majority}$)
        iii) Select mixed instance I_{mix} from $Bank$: $I_{mix} = Bank[c_{mixed}]$
      else
        The current c_i is the minority base class and the responding instance I_{base} is the minority base instance
        i) $V_{minority} = M(c_i) \in R^{1 \times N_c}$
        ii) Weight sampling algorithm: $c_{mixed} = torch.multinomial(V_{minority})$
        iii) Select mixed instance I_{mix} from $Bank$: $I_{mix} = Bank[c_{mixed}]$
      end if
    end for
  end for
  ***Step 3*:** Mixup augmentation: Obtain a class-rebalanced image $\hat{I}$ by Eq. (5)
end while
Output: a class-rebalanced image $\hat{I}$

all samples equally, our method selectively samples based on whether they belong to the minority or majority base classes. Second, we create a confusion matrix and use a weight sampling algorithm to select the mixed instance I_{mix} that best matches a base instance I_{base} in the source image. This focused approach is theoretically more robust than blind sampling, as it focuses on correcting the root causes of class confusion. Third, incorporating Mixup within this confusion-matrix-guided instance mixing strategy expands the minority class features into the feature space of the majority class, ensuring that the augmented samples explicitly address the model's weaknesses in distinguishing the challenging class.

3.3. Self-supervised learning based ISP degradation

Self-supervised learning based ISP degradation contains a day-to-night transformation module (ISP-DTM) and a self-supervised learning strategy.

3.3.1. ISP-DTM

As shown in Fig. 5, the ISP-DTM consists of three steps: i) Invert Processing step, ii) Noise Modeling step and iii) ISP Pipeline step. Specifically, the Invert Processing step contains (1) Invert Tone Mapping, (2) Invert Gamma Correction, (3) Color Transformation $y_{s \to c}$ and (4) Invert White Balance. Based on this step, the realistic RAW format (*i.e.* cRGB) images are generated and we denote (1), (2), (3), (4) together as T_{invert}. Then, considering the physical noise of camera sensors, we model two common noises in camera (*i.e.* 'shot' and 'read' noise) in the Noise Modeling step and we denote the output of this step as y_{nm}. Finally, the cRGB with shot and read noises is restored back to sRGB by the ISP Pipeline step for dark object detection. As shown in the green area of Fig. 5, the ISP Pipeline step contains (5) Signal Quantization, (6) White Balance, (7) color transformation $y_{c \to s}$ and (8) Gamma Correction, and we define (5), (6), (7), (8) as T_{ISP}. Next, we will describe each process of ISP-DTM in detail:

White Balance. The human eyes have color constancy, *i.e.* human perception of the color tends to be stable under the change of illumination condition. However, the camera sensor does not have this characteristic, resulting in color shift, and the white balance algorithm is proposed to correct this color deviation. Specifically, this algorithm balances the channel gain of red g_r and blue g_b to make images appear to be lit under the neutral illumination [7].

Color Transformation. Because the data format of standard color space (sRGB) does not match the camera internal color space (cRGB), we use a 3×3 color correction matrix T_{ccm} to achieve this color transformation: $y_{c \to s} = T_{ccm} \cdot I_{cRGB}$ and $y_{s \to c} = T_{ccm}^{-1} \cdot I_{sRGB}$, where $y_{c \to s}$ denotes the color transformation from the camera internal color space (I_{cRGB}) to the final standard color space (sRGB) and $y_{s \to c}$ denotes the invert process.

Gamma Correction. The purpose of gamma correction is to adjust the overall light and dark values of images. If the original image collected by the camera sensor is not processed by the gamma correction, it will adversely affect the results of dark object detection due to

problems of illumination and shadows. Gamma correction controls the overall brightness of the image through two parameters: $I_{out} = \alpha I_{in}^{1/\gamma}$, where I_{in} and I_{out} denote input and output images, α and γ are used to adjust the shape of the gamma correction curve. In this paper, γ is sampled from a uniform distribution $\gamma \sim U(2, 3.5)$ and α is set to 1. The invert process of gamma correction is to replace $1/\gamma$ with γ: $I_{out} = \alpha I_{in}^{\gamma}$.

Tone Mapping. High dynamic range images in real scenes require tone mapping operation to suit the dynamic range of camera sensors [29]. Here, we simplify the tone mapping to a simple 'smooth-step' curve as in Eq. (6) and its invert process is in Eq. (7) :

$$F_{tm}(x) = 3x^2 - 2x^3 \tag{6}$$

$$F_{tm}^{-1}(y) = \frac{1}{2} - \sin(\frac{\sin^{-1}(1-2y)}{3}) \tag{7}$$

Noise Modeling. Noises of camera sensors primarily come from two sources: 'shot' noise leads to fluctuations in the gray value of the images and 'read' noise generated by the electronics in the readout cameras. Mathematically, shot noise is a Poisson random variable whose mean is the light intensity (*i.e.* parameter k in Eqs. (8) and (11)) and read noise is a Gaussian random variable with zero mean and fixed variance [30]. We model both of them as x_{noise}:

$$x_{noise} \sim N(\mu = 0, \sigma^2 = k \cdot I \cdot \lambda_{shot} + \lambda_{read}) \tag{8}$$

where I is the output image from the Invert Processing step, λ_{shot} and λ_{read} are digital and analog gains of camera sensors, which could be sampled from the joint distribution of different shot/read noise parameter pairs in RAW images [30]. The details of the sampling process are:

$$\log \lambda_{shot} \sim U(a = \log(0.0001), b = \log(0.012)) \tag{9}$$

$$\log \lambda_{read} \sim N(\mu = 2.18 \log \lambda_{shot}, \sigma = 0.26) \tag{10}$$

Moreover, parameter k in Eq. (8) is the light intensity (between 0.01 and 1.0), and it follows a truncated Gaussian distribution [7]:

$$k \sim N(\mu = 0.1, \sigma = 0.08) \tag{11}$$

Finally, the outputs y_{nm} of the Noise Modeling step can be formulated as: $y_{nm} = k \cdot I + x_{noise}$, which is then sent to the ISP Pipeline step for subsequent transformation processing.

Signal Quantization. The first process in the ISP Pipeline step is to quantize y_{nm} by an analog-to-digital converter (ADC). In this paper, we simulate this process as:

$$\hat{y} \sim (-\frac{1}{2B}, \frac{1}{2B}) \tag{12}$$

$$y_{quant} = y_{nm} + \hat{y} \tag{13}$$

where B is randomly selected from 12, 14, and 16 as in [7].

Moreover, during the process of ISP-DTM, we calculate four parameters, *i.e.* light intensity k in Eq. (11), $1/\gamma$ in Invert Gamma Correction, channel gain $1/g_r$ and $1/g_b$ in Invert White Balance, which are used as the ground truth in the following self-supervised learning strategy. We note that all four parameters (k, $1/\gamma$, $1/g_r$, $1/g_b$) are global rather than spatially varying. Furthermore, to improve reproducibility, we have open-sourced our code.

In summary, for a daytime image I, we obtain a low-light image I_l and four ground truths $p_i (i = 1, 2, 3, 4)$ by ISP-DTM, and the whole process can be expressed:

$$I_l + p_i = T_{ISP}\left[T_{invert}(I) + y_{nm}\right] \tag{14}$$

3.3.2. Self-supervised learning strategy

After obtaining low-light images, we compose low-light images and daytime images into image pairs. Then, we utilize an Encoder-Decoder to encode the pair of images into high-level features by a weight-shared Encoder, and to decode four parameters $\tilde{p}_i$ ($\tilde{k}$, $1/\tilde{\gamma}$, $1/\tilde{g}_r$, $1/\tilde{g}_b$) as the degradation predictions. The loss of self-supervised learning strategy L_{self} is a L_1 loss:

$$L_{self} = \frac{1}{4}\sum_{i=1}^{4} L_1(p_i, \tilde{p}_i) \tag{15}$$

where p and $\tilde{p}$ denote the ground truth and prediction of the parameters k, $1/\gamma$, $1/g_r$, $1/g_b$, respectively. The weights of k, $1/\gamma$, $1/g_r$, $1/g_b$ are set to 5:1:1:1 in our implementation.

3.4. Disentanglement regularization (DR)

As shown in Fig. 3, the encoder in our model has two functions: i) encode the pair of daytime and nighttime images for learning the parameters of ISP, ii) extract the feature for training the detector. The task-specific decoders are used to output two different aspects, *i.e.* ISP-related parameters for self-supervised learning and bounding boxes and classes of object detection. However, this multi-task learning framework may cause the problem of conflicting gradients.

To overcome this issue, we propose a regularization to disentangle these two tasks (*i.e.* ISP degradation and object detection) in the training process. Specifically, as shown in the bottom part of Fig. 3, the red arrow g_1 and blue arrow g_2 denote gradient vectors of the task of object detection and self-supervised learning based ISP degradation respectively. For different tasks, we minimize cosine similarity by pushing the g_1 or g_2 close to the dotted line arrow under the supervision of DR loss L_{DR}. For the same tasks, we make the gradient vectors as coincident as possible. Mathematically, DR can be expressed as:

$$L_{DR} = \omega_1\left|\cos(g_1, g_2)\right| + \omega_2(\left|1 - \cos(g_1, g_1)\right|) + \omega_3(\left|1 - \cos(g_2, g_2)\right|) \tag{16}$$

where ω_1, ω_2 and ω_3 are the parameters to balance these terms. We set $\omega_1 = 5$ and $\omega_2 = \omega_3 = 0.5$.

3.5. Dynamic IoU-threshold adjustment strategy

For the teacher-student architecture based unsupervised domain adaptive methods, pseudo-labels $Y_p = \{p_b^t, p_c^t\}$ are generated by the teacher model and filtered by a static threshold τ_{cls} which is unchanged during the training [9,10]. However, these methods fail to take into account that the classification confidences would continue to change across categories and iterations [3]. As a result, many potential True-Positive pseudo-labels, *i.e.* correctly predicted by the teacher network, are excessively filtered out by a high static threshold τ_{cls}, which is usually set to 0.8.

To overcome this issue, our goal is to adjust the threshold τ_{cls} dynamically to find True-Positive pseudo-labels that have been filtered out. We propose a Dynamic IoU-threshold Adjustment (DIA) strategy using the IoU that is calculated by bounding-box predictions of the teacher and student models to update the threshold during training. Specifically, as shown in Fig. 6(a), just like traditional teacher-student architecture based UDA methods, the scores of teacher classification prediction $S_c^t \in R^{100\times1}$ from the teacher model are filtered by a threshold τ_{cls} (the red dotted line in Fig. 6(a), the value is initialized to 0.8) firstly, which is formulated as: $\hat{S}_c^t = where\ S_c^t > \tau_{cls}$, where the type of $\hat{S}_c^t \in R^{100\times1}$ is *Bool*. Through the previous analysis, we know True-Positive pseudo-labels that are filtered out mistakenly may exist in $\hat{S}_c^t = False$ (the purple point under the red dotted line in Fig. 6(a)).

Then, the IoU of bounding-box predictions can be calculated by: $M_{iou}(t, s) = F_{iou}(p_b^t, p_b^s)$, where $p_b^t \in R^{100\times4}$ and $p_b^s \in R^{100\times4}$ denote bounding-box predictions of teacher and student models, respectively. $F_{iou}(\cdot)$ is a function to calculate IoU, and IoU matrix $M_{iou}(t, s) \in R^{100\times100}$ is the results. Furthermore, in $M_{iou}(t, s)$, only those values exceeding an IoU threshold τ_{iou} are retained according to Eq. (17):

$$\hat{M}_{iou}(t, s) = where\ M_{iou}(t, s) > \tau_{iou} \tag{17}$$

where τ_{iou} denotes the IoU threshold that is set to 0.8 as default, and the type of $\hat{M}_{iou}(t, s)$ is *Bool*. As shown in Fig. 6(b), the blue square areas of $\hat{M}_{iou}(t_i, s_i)$ are $True$ for mining True-Positive ones and yellow square areas of $\hat{M}_{iou}(t_i, s_i)$ are $False$, which would be discarded.

After that, we define that the filtered pseudo-labels are considered True-Positive when the following conditions are met: i) $\hat{S}_c^t = False$,

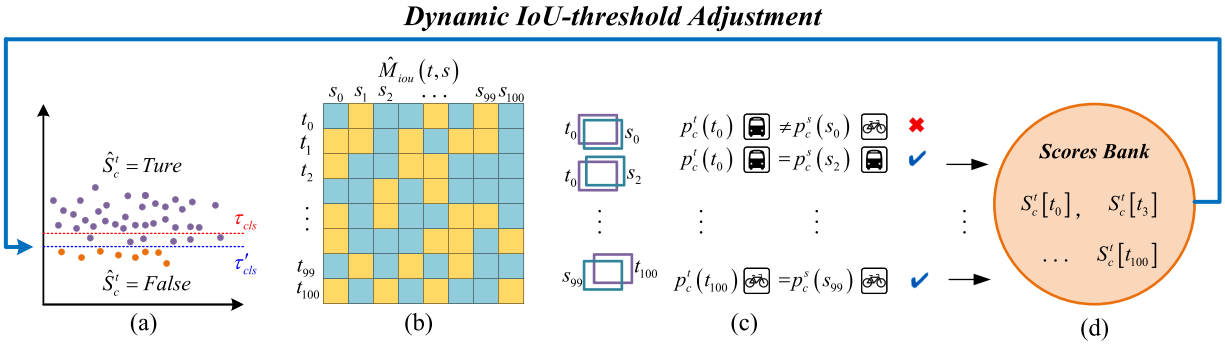

Fig. 6. The architecture of our Dynamic IoU-threshold Adjustment (DIA). **(a)** Traditional methods obtain pseudo-labels through a static threshold τ_{cls}, but some True-Positive samples may be filtered out (*e.g.* purple points under the red dotted line). τ'_{cls} is the updated threshold obtained by our Dynamic IoU-theshold Adjustment. **(b)** A *Bool* type of IoU matrix $\hat{M}_{iou}(t,s)$, which horizontal axis t_i represents the bounding box of teacher, and the vertical axis s_i represents the bounding box of student, respectively. The blue areas in the matrix denote the IoU of the teacher and student bounding boxes are greater than a threshold τ_{iou}, and the yellow areas denote that IoU is less than the threshold. **(c)** In each part where IoU is greater than a threshold τ_{iou} (*i.e.* blue areas), only the teacher and student models predicting the same class would be saved, *e.g.* the class corresponding to bounding boxes t_0 and s_2 are both 'Bus', or the class corresponding to bounding boxes t_{100} and s_{99} are both 'Bicycle'. The teacher and student models predicting different classes would be discarded. *e.g.* the class corresponding to bounding boxes t_0 is 'Bus' while the class corresponding to bounding boxes s_0 is 'Bicycle'. **(d)** For t_i saved classes in the step (c), which the corresponding pseudo-labels are considered to be True-Positive samples and we store these scores of this t_i in a Scores Bank. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

which means True-Positive pseudo-labels that are filtered out mistakenly may exist in $\hat{S}_c^t = False$ and our goal is to recover them to $True$. ii) The IoU of teacher and student bounding-box predictions needs to exceed an IoU threshold τ_{iou}, *i.e.* meets Eq. (17). iii) As illustrated in Fig. 6(c), we iterate the elements in the IoU matrix $\hat{M}_{iou}(t_i, s_i)$, where the teacher and student models predict the same classes to be saved, *i.e.* $p_c^t(t_i) = p_c^s(s_i)$. As a result, the pseudo-labels corresponding to this t_i are considered to be True-Positive samples. We recover them by $\hat{S}_c^t[t_i] = True$ and store the scores of these t_i in the Scores Bank, *i.e.* $Bank_c = S_c^t[t_i]$, for updating the threshold τ_{cls} in subsequent iterations.

Finally, we update the threshold in the following three steps: i) We first sort the scores in *Scores Bank* from high to low and take the first 50% to calculate their mean μ' and standard deviation σ'. ii) Inspired by [20], an auxiliary threshold update step is introduced: $\tau_{step} = \mu' - 2 \times \sigma'$. iii) The threshold is updated by the following formula:

$$\tau'_{cls} = \tau_{cls} + \phi(\tau_{step} - \tau_{cls}) \tag{18}$$

where τ'_{cls} is the updated threshold and ϕ is an adjustment parameter set by 0.05 as default.

3.6. Total loss

The total loss function includes L_{self} loss that makes the encoder capture the intrinsic visual information, L_{DR} pushes two gradient vectors at different tasks as orthogonal as possible, and the original detection losses L_{det} in the Teacher-Student architecture:

$$L_{total} = \beta L_{self} + L_{DR} + L_{det} \tag{19}$$

where β is the weight of self-supervised learning loss.

4. Experiments

In this section, we conduct experiments to validate our proposed ISP Dynamic Teacher on BDD100k [12] and SHIFT [13] datasets. First, we give a brief introduction of the used datasets, metrics, and implementation details. Then, we show the main results of our method on the two datasets and we compare our proposed method with our conference version method ISP-Teacher and some SOTA methods. Moreover, some ablation studies are conducted to verify the effectiveness of each component and some hyper-parameter sensitivities in our method. Finally, we show some qualitative results to further validate our proposed method.

4.1. Datasets and metrics

BDD100k is a widely used autonomous driving dataset [12], which includes 10 common classes and covers various weather scenarios. Following [11], we split the BDD100k dataset into two parts using labels 'day' and 'night'. Specifically, daytime images and nighttime images are used as source and target data for training respectively, and only nighttime images in the validation dataset are used for validation. After splitting, there are 36,728 daytime images and 32,998 nighttime images in the training set and 4707 nighttime images in the validation set.

SHIFT is also an autonomous driving dataset [13], and it includes discrete shifts (*e.g.* urban, village, and rural) and continuous shifts (*e.g.* daytime to night) in cloudiness, rain, and fog weather. SHIFT dataset has the same 6 classes as BDD100k dataset with bounding box annotations. Similar to BDD100k, we also split it into 19,452 daytime images and 8497 nighttime images for training and 1200 nighttime images for validation.

ACDC [31] is a large-scale dataset designed to evaluate robust perception under challenging conditions. It contains 406 images for validation, which covers night, snow, fog, and rain adverse conditions. In this paper, we specifically use night subset ACDC (night) for cross dataset testing.

Cityscapes and Foggy Cityscapes. Cityscapes [32] is a widely used street scene dataset, which contains 2975 training images and 500 validation images. Foggy Cityscapes [33] introduces heavy foggy noise and simulates a synthesized dataset.

As for the metrics, following [11], we adopt AP_{50} (IoU@0.5), AP_S (small-sized object), AP_M (medium-sized object) and AP_L (large-sized object) to evaluate our proposed method.

4.2. Implementation details

Following previous teacher-student architecture based domain adaptive methods, we use Faster-RCNN [5] with ResNet50 [34] as our baseline detector. SGD is used as the optimizer with a base learning rate of 0.01, and the momentum is set to 0.9. Loss hyper parameter $\lambda = 0.3$ and $\beta = 1$, and the rate smooth coefficient parameter of EMA is set to 0.9996. The batch size is set to 4, which includes 2 daytime images in the source domain and 2 nighttime images in the target domain, and all images are proportionally scaled to a minimum side of 600.

In the DIA strategy, τ_{cls} is initialized to 0.8, which changes dynamically during the training. The value of τ_{iou} is fixed at 0.8 and the

Table 1
Main results of our proposed method on BDD100k dataset. We show the AP_S, AP_L and AP_{50} of each class, along with the improvements of our method over all baseline approaches. The classes are organized by semantic categories into Human-related (Pedestrian, Rider), Vehicle-related (Car, Truck, Bus, Motorcycle, Bicycle), and Traffic-related (Traffic Light, Traffic Sign). Source (Lower-Bound) denotes the results of training Faster-RCNN with only daytime images and testing on nighttime images, and Oracle (Upper-Bound) denotes the results of training Faster-RCNN on nighttime images with ground-truth and testing on nighttime images. Note that **bold** and underline results represent the best and second-best performance for each class, respectively.

Method	AP_{50}	AP_S	AP_L	Human-related		Vehicle-related					Traffic-related	
				Ped.	Rid.	Car	Tru.	Bus	Mot.	Bic.	T-L.	T-S.
Source	41.1	6.8	35.7	50.0	28.9	66.6	47.8	47.5	32.8	39.5	41.0	56.5
Oracle	46.2	8.8	40.2	52.1	35.0	73.6	53.5	54.8	36.0	41.8	52.2	63.3
UMT [35]	36.2	5.1	34.5	46.5	26.1	46.8	44.0	46.3	28.2	40.2	31.6	52.7
	+13.4	+4.2	+11.7	+11.4	+14.3	+26.8	+13.5	+10.1	+14.7	+10.1	+19.4	+13.3
TDD [4]	34.6	6.3	30.7	43.1	20.7	68.4	33.3	35.6	16.5	25.9	43.1	59.5
	+15.0	+3.0	+15.5	+14.8	+19.7	+5.2	+24.2	+20.8	+26.4	+24.4	+7.9	+6.5
AT [10]	38.5	6.0	32.9	42.3	30.4	60.8	48.9	52.1	34.5	42.7	29.1	43.9
	+11.1	+3.3	+13.3	+15.6	+10.0	+12.8	+8.6	+4.3	+8.4	+7.6	+21.9	+22.1
2PCNet [11]	46.4	9.1	41.7	54.4	30.8	73.1	53.8	55.2	37.5	44.5	49.4	65.2
	+3.2	+0.2	+4.5	+3.5	+9.6	+0.5	+3.7	+1.2	+5.4	+5.8	+1.6	-0.3
De-Teacher [21]	46.2	-	-	52.8	35.7	63.2	54.8	50.2	**45.7**	46.0	45.8	**68.0**
	+3.4	-	-	+5.1	+4.7	+10.4	+2.7	+6.2	-2.8	+3.7	+5.2	-2.0
Cos [20]	49.4	**10.2**	45.9	57.1	**41.3**	**73.9**	54.7	55.6	43.5	50.2	50.0	67.8
	+0.2	-0.9	+0.3	+0.8	-0.9	-0.3	+2.8	+0.8	-0.6	+0.1	+1.0	-1.8
ISP-Teacher [15]	48.8	9.2	45.7	57.8	39.4	72.9	54.6	55.9	43.8	48.1	49.6	66.3
	+0.8	+0.1	+0.5	+0.1	+1.0	+0.7	+2.9	+0.5	-1.1	+2.2	+1.4	-0.3
Ours	**49.6**	9.3	**46.2**	**57.9**	40.4	73.6	**57.5**	**56.4**	42.9	**50.3**	**51.0**	66.0

adjustment parameter ϕ is set to 0.05. For the burn-up stage, we train the student network in a supervised manner in the source domain for 50k and 40k iterations for BDD100k and SHIFT datasets, respectively. The total iterations on BDD100k and SHIFT are 110k and 90k iterations. Our method is implemented on 4 RTX6000 GPUs.

4.3. Main results

In order to verify the effectiveness of ISP Dynamic Teacher for dark object detection, we compare our method with some SOTA methods, *i.e.* UMT [35], TDD [4], AT [10], 2PCNet [11], ISP-Teacher [15] and Cos [20]. It would be emphasized that the proposed method ISP Dynamic Teacher in this paper is an extended version of ISP-Teacher [15] and Cos [20] is the most recent object detection method specially designed for low-light images and it achieves SOTA performance on the BDD100k dataset.

4.3.1. Experiments on BDD100k

As shown in Table 1, the previous teacher-student architecture methods achieve terrible results in night scenes and are even much lower than the Lower-Bound. For example, TDD [4] and UMT [35] only get 34.6% AP and 36.2% AP in night scenes. And although AT [10] improves the performance (38.5% AP) with the design of domain adversarial learning and weak-strong data augmentation, it is still far behind (-2.6%) the baseline detector Faster-RCNN (*i.e.* 'Source', 41.1% AP). Through the four innovations in this paper, our ISP Dynamic Teacher outperforms all the teacher-student architecture methods by a large margin (over 3% in AP). Specifically, compared to the method of only training on daytime source images and testing on nighttime ('Source' in the first row), our method brings a +8.5% AP improvement (*i.e.* 49.6% vs. 41.1%). Furthermore, our unsupervised approach even outperforms the supervised method 'Oracle' (the second row) that trains on nighttime images with annotations by +3.4% in AP.

Compared with our conference version ISP-Teacher [15], the performance of our ISP Dynamic Teacher (an extended version method) is greatly improved thanks to the design of the class-rebalanced module and dynamic IoU-threshold adjustment strategy. As shown in Table 1, we can see that our extended version method brings a +0.8% improvement in AP (from 48.8% to 49.6%) and it improves performance in seven out of nine categories, where 'Traffic Sign' is only 0.3% lower. Particularly, the most huge improvement is on the 'Truck' class (from 54.6% to 57.5%, +2.9%). Moreover, there is a notable improvement on the 'Bicycle' class (from 48.1% to 50.3%, +2.2%). Furthermore, ISP Dynamic Teacher achieves a large margin improvement (from 49.6% to 51.0%, +1.4%) on the 'Traffic Light' class and an impressive result of 40.4% is received on the 'Rider' class, surpassing the ISP-Teacher method by +1.0% absolutely. Finally, our method also achieves satisfactory results on the 'Car', 'Bus' and 'Pedestrian' classes, which obtain +0.7%, +0.5% and +0.1% improvements, respectively. The above comparison clearly proves the effectiveness of our proposed method on dark object detection.

For the dark object detection methods, ISP Dynamic Teacher achieves the SOTA performance. From the last three rows of the Table 1, we can see that our proposed method achieves a new SOTA (49.6% AP) in dark object detection. Specifically, ISP Dynamic Teacher obtains a +2.8% AP improvement (57.5% *vs.* 54.7%), where is the most notable improvement on the 'Truck' class. Furthermore, our method outperforms the class of 'Traffic Light', 'Pedestrian', and 'Bus' by 1.0% (from 50.0% to 51.0%), 0.8% (from 57.1% to 57.9%) and 0.8% (from 55.6% to 56.4%) in terms of AP, respectively. Moreover, compared to the 'Bicycle' (50.3% *vs.* 50.2%) and 'Car' (73.6% *vs.* 73.9%) classes, we can also get similar results. However, the performance of ISP Dynamic Teacher on some classes is not satisfactory, which is 1.8% and 0.9% lower on the 'Traffic Sign' and 'Rider' class, respectively. We think this will also be the direction that we optimize the algorithm in the future.

4.3.2. Experiments on SHIFT

To further verify the effectiveness of our proposed method, we conduct experiments on the SHIFT dataset and the results are shown in Table 2. We can see that other teacher-student architecture based methods also perform worse than Lower-Bound (*e.g.* UMT [35] and AT [10] are 10.5% and 2.7% lower than Lower-Bound, respectively), and our ISP Dynamic Teacher outperforms all other methods as well.

Note that our conference version method ISP-Teacher [15] still is the SOTA performance so far. Moreover, compared with ISP-Teacher, our extended version method ISP Dynamic Teacher achieves +0.4% improvement in AP (from 52.4% to 52.8%). From Table 2, we know that the reason is mainly due to the improvement on the 'Bus', 'Motorcycle'

Table 2
Main results of our proposed method on SHIFT dataset. Source (Lower-Bound) denotes the results of training Faster-RCNN with only daytime images and testing on nighttime images, and Oracle (Upper-Bound) denotes the results of training Faster-RCNN on nighttime images with ground-truth and testing on nighttime images. Note that **bold** and underline results represent the best and second-best performance within each class.

Method	AP_{50}	AP_S	P_L	Human-related	Vehicle-related				
				Pedestrian	Car	Truck	Bus	Motorcycle	Bicycle
Source	41.6	6.4	65.0	40.4	44.5	49.9	53.7	14.3	46.7
Oracle	47.0	9.6	71.3	49.7	51.5	56.0	53.6	19.2	52.4
UMT [35]	31.1	3.1	45.9	7.7	47.5	18.4	46.8	16.6	49.2
	+21.7	+10.1	+34.5	+44.0	+11.3	+40.1	+17.3	+8.0	+9.6
AT [10]	38.9	5.2	61.5	25.8	33.0	54.7	49.5	20.7	52.3
	+13.9	+8.0	+18.9	+25.9	+25.8	+3.8	+14.6	+3.9	+6.5
2PCNet [11]	49.1	10.6	76.0	51.4	54.6	54.8	56.6	23.9	54.2
	+3.7	+2.6	+4.4	+0.3	+4.2	+3.7	+7.5	+0.7	+4.6
De-Teacher [21]	52.5	-	-	**52.3**	57.7	58.4	62.3	**27.1**	56.9
	+0.3	-	-	-0.6	+1.1	+0.1	+1.8	-2.5	+1.9
Cos [20]	51.0	12.3	78.9	50.9	56.0	57.2	**64.5**	22.2	55.5
	+1.8	+0.9	+1.5	+0.8	+2.8	+1.3	-0.4	+2.4	+3.3
ISP-Teacher [15]	52.4	13.1	80.0	51.6	**59.1**	**58.7**	62.3	24.1	58.3
	+0.4	+0.1	+0.4	+0.1	-0.3	-0.2	+1.8	+0.5	+0.5
Ours	**52.8**	**13.2**	**80.4**	51.7	58.8	58.5	64.1	24.6	**58.8**

and 'Bicycle' classes, *i.e.* minority classes on SHIFT dataset. Specifically, our ISP Dynamic Teacher obtains a huge improvement on the 'Bus' class (from 62.3% to 64.1%, +1.8%) and +0.5% improvement is achieved on both class 'Motorcycle' and class 'Bicycle' (24.1% *vs.* 54.6%, 58.3% *vs.* 58.8%). These results further prove the effectiveness of our new contributions (*i.e.* class-rebalanced module and dynamic IoU-threshold adjustment strategy) added to the extended version.

Compared with the dark object detection method, our ISP Dynamic Teacher outperforms a night-specific algorithm 2PCNet [11] by 3.7% in AP (from 49.1% to 52.8%). Furthermore, although Cos [20] achieved good results on the BDD100k dataset, it performed poorly on the SHIFT dataset, even worse than our conference version method ISP-Teacher [15] (*i.e.* 51.0% *vs.* 52.4%). From the last three rows of Table 2, our ISP Dynamic Teacher achieves a new SOTA performance *i.e.* +52.8% in AP and obtains a +1.8% improvement compared with Cos [20] (51.0% *vs.* 52.8%). Specifically, ISP Dynamic Teacher achieves 58.5% AP on the 'Bicycle' class, surpassing Cos [20] by a large margin, *i.e.* +3.3% in AP (from 55.5% to 58.8%). Furthermore, our method obtains a huge improvement on the 'Car' class (from 56.0% to 58.8%, +2.8%), and a notable improvement (+2.4%) on the 'Motorcycle' is obtained compared with Cos [20] (22.2% *vs.* 24.6%). The above experiments further demonstrate the effectiveness of our proposed method in the dark object detection task.

4.3.3. Experiments on minority classes

Our method yields substantial improvements in minority classes such as 'Truck', 'Bus', 'Motorcycle', 'Bicycle'. For instance, as shown in Table 1, our approach improves the AP of Motorcycle by +8.6% and +24.2% over AT [10] and TDD [4] baselines, respectively. This confirms the effectiveness of our CRM in alleviating the class imbalance problem. The combination of self-supervised Learning Based ISP Degradation (ISP-DTM) and the Dynamic IoU-threshold Adjustment (DIA) strategy significantly boosts the detection performance of small objects AP_S. Specifically, AP_S increases from 6.3% to 9.3% on the BDD100k dataset (Table 1) and from 5.2% to 13.2% on SHIFT dataset (Table 2) compared to the AT [10] baseline. These results further demonstrate that the ISP-DTM preserves texture and structural details for small objects under low-light conditions, while DIA reduces false negatives for high-quality small pseudo-labels.

4.3.4. Cross datasets and different weather conditions experiments

i) **Cross datasets experiments.** We use daytime images from the BDD100k dataset as the source domain and nighttime images from the ACDC dataset as the target domain, since the two datasets share the same set of classes. As shown on the left side of Table 3, 2PCNet [11] and Cos [20] show limited performance under the cross dataset setting, achieving only 11.9% and 13.7% in AP. Our method achieves the best performance on all object scales, surpassing previous SOTA method Cos [20] by +3.3% in AP, indicating that the proposed learnable ISP-based degradation module and dynamic class-rebalanced and IoU-threshold strategies effectively enhance feature representation under cross nighttime dataset conditions.

ii) **Different weather conditions experiments.** We conduct experiments on BDD100k (day → rainy) and Cityscapes → Foggy Cityscapes setting to evaluate generalization under rainy and foggy conditions. As shown on the middle and right sides of Table 3, our method consistently outperforms strong UDA baselines under rainy and foggy conditions in terms of AP_{50} (53.1% in rainy and 43.8% in foggy). Furthermore, the strong AP_S results across both domains (*i.e.* 12.0% in rainy and 3.3% in foggy conditions) indicate that the CRM and DIA modules of our method are essential for small object detection. The gains under rainy and foggy conditions demonstrate that our method is not limited to nighttime illumination changes, but generalizes to other weather-induced domain shifts.

4.4. Ablation study

To validate the effectiveness of each component in our proposed method, we conduct some ablation experiments on the BDD100k dataset. Moreover, some analyzes of the hyper-parameters are also shown in this section.

4.4.1. Effectiveness of each component

i) Self-supervised Learning Based ISP Degradation. We compare our self-supervised learning based ISP degradation (Self., the third row of Table 4) with other methods: (1) traditional non-learnable data augmentations like randomly color jittering, gray scaling, Gaussian blurring (the first row of Table 4), and (2) nighttime specific augmentations NightAug in 2PCNet [11] (NA., the second row in Table 4). As shown in Table 4, we see that the method of traditional non-learnable data augmentation conducted on daytime images obtains poor performance (42.2% AP) in low-light conditions. The NightAug method (in the second row) is a nighttime specific data augmentation that aims to reduce the bias between daytime and nighttime images, and it brings +3.6% AP improvement compared to the non-learnable data augmentation method (42.2% *vs.* 45.8%). Our proposed self-supervised learning

Table 3
Left: cross nighttime dataset experiments from BDD100k → ACDC (night) setting. Middle and Right: different weather conditions from BDD100k (day → rainy) and Cityscapes → Foggy Cityscapes, respectively.

Method	BDD100k → ACDC (night)				BDD100k (day → rainy)				Cityscapes → Foggy Cityscapes			
	AP_{50}	AP_S	AP_M	AP_L	AP_{50}	AP_S	AP_M	AP_L	AP_{50}	AP_S	AP_M	AP_L
2PCNet [11]	11.9	1.7	8.2	12.9	48.4	10.9	34.1	40.0	43.0	3.2	20.2	47.6
Cos [20]	13.7	1.8	12.4	19.4	49.0	10.1	34.4	48.5	41.8	3.2	19.7	44.6
ISP-Teacher [15]	16.3	1.8	13.0	21.3	51.9	11.6	35.3	50.9	43.3	3.1	**20.7**	47.6
Ours	**17.0**	**1.9**	**13.1**	**23.3**	**53.1**	**12.0**	**36.7**	**51.7**	**43.8**	**3.3**	20.4	**47.9**

Table 4
Left: ablation study of each component in our ISP Dynamic Teacher on BDD100k dataset. Top right and bottom right: the influence of different weights ϕ in the dynamic IoU-threshold adjustment strategy and β in the self-supervised learning based ISP degradation loss. 'NA.' denotes a non-learnable nighttime specific augmentation in 2PCNet, *i.e.* NightAug. 'Self.' and 'DR' denote the self-supervised learning based ISP degradation and disentanglement regularization. CRM and DIA denote Class-rebalanced Module and Dynamic IoU-threshold Adjustment, which are the new contributions proposed in the extended version ISP Dynamic Teacher. AP_S, AP_M and AP_L denote the AP on small-sized, medium-sized and large-sized objects, respectively.

NA.	Self.	DR	CRM	DIA	AP_{50}	AP_S	AP_M	AP_L	ϕ	AP_{50}	AP_S	AP_M	AP_L
-	-	-	-	-	42.2	7.9	23.0	38.8	w/o	49.2	9.2	27.0	**46.8**
✓	-	-	-	-	45.8	8.6	25.7	42.2	0.03	49.1	9.0	27.3	46.5
-	✓	-	-	-	48.5	9.2	27.1	45.2	0.05	**49.6**	**9.3**	**27.3**	46.2
-	-	-	✓	-	43.4	8.3	24.5	39.7	0.07	48.8	9.2	27.1	46.5
-	-	-	-	✓	42.9	7.9	23.9	39.1	β	AP_{50}	AP_S	AP_M	AP_L
-	✓	✓	-	-	48.8	9.2	27.2	45.7	1	**49.6**	**9.3**	**27.3**	**46.2**
-	✓	✓	✓	-	49.2	9.2	27.0	46.8	2	49.2	9.2	27.2	46.0
-	✓	✓	✓	✓	49.6	9.3	27.3	46.2	5	46.8	8.1	25.9	45.2

based ISP degradation not only addresses the domain bias of the student network but also explores how to learn intrinsic visual information of dark images, which achieves a huge improvement in AP (from 42.2% to 48.5%, +6.3%). This module mitigates domain bias by modeling the camera imaging process, as demonstrated by its effectiveness in object detection on low-light images.

ii) Disentanglement Regularization. As shown in the sixth row of Table 4, when adding disentanglement regularization (DR) in our framework, the detection performance can further improve to 48.8% from 48.5%. This is due to Disentanglement Regularization explicitly minimizes the cosine similarity between gradients of object detection and self-supervised learning based ISP degradation, effectively mitigating gradient conflict. The above are the contributions of ISP-Teacher [15], based on it, we add two new contributions in the following.

iii) Class-rebalanced Module. From Table 4, we can see that after adding the class-rebalanced module, the detection performance increases from 48.8% to 49.2% (+0.4% in AP), where it bring a huge improvement (+1.1%) on large-sized object, *i.e.* AP_L from 45.7% to 46.8%. At the same time, there is basically no impact on small-sized and medium-sized objects.

In addition, using the CRM alone (row 4 in Table 4) also brings a noticeable improvement (42.2% *vs.* 43.4% in AP). These comparisons prove that Class-Rebalanced Module addresses the class imbalance problem by redistributing class samples based on the confusion matrix and a weight sampling algorithm, which improves the accuracy of object detection tasks.

iv) Dynamic IoU-threshold Adjustment. Finally, when using the dynamic IoU-threshold adjustment strategy, as shown in the last row of Table 4, the performance further improved to a new SOTA result (+49.6% in AP), and it also brings +0.4% improvement in AP. We found that the improvement is mainly concentrated on small-sized and medium-sized objects. Moreover, even when used alone, the DIA still leads to performance improvement (from 42.2% to 42.9% in AP). The reason is that the DIA strategy enhances pseudo-label quality by recovering true positives that are mistakenly filtered out by static thresholds (especially small objects).

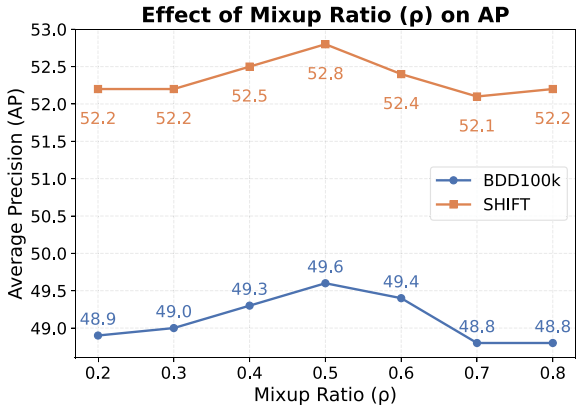

Fig. 7. Analysis of mixup ratio ρ in Mixup augmentation. As the mixup ratio increases, the performance shows similar trends on both BDD100k and SHIFT dataset, achieving the best performance at $\rho = 0.5$.

4.4.2. *Analysis of hyper-parameter*

i) Analysis of weights ϕ of the dynamic IoU-threshold adjustment strategy. From Eq. (18) we know that $\phi = 0$ means dynamic IoU-threshold adjustment (DIA) strategy is not used. As shown in Table 4, the result without using the DIA strategy is 49.2% in AP. When ϕ increases to 0.03, the result drops slightly (from 49.2% to 49.1%), and when ϕ continues to increase to 0.05, our method achieves the SOTA result (49.6% in AP). At the same time, it also obtains the best results on small-sized objects (9.3% in AP_S) and medium-sized objects (27.3% in AP_M). However, as ϕ continues to increase, the result begins to decline and is even 0.4% lower than the result without using the DIA strategy (48.8% *vs.* 49.2%). From the above results, we conclude that the hyper-parameter ϕ is sensitive to the final result, so we set $\phi = 0.05$ as default.

ii) Analysis of weights β in the self-supervised learning based ISP degradation loss. From Eq. (19), we add the self-supervised learn-

Fig. 8. Examples of detection results on BDD100k dataset. For each row of figure, (a) and (b) are dark scenes in clear weather, (c) is the dark scene in rainy weather, (d) contains minority class 'Bus', (e) and (f) are small instances of different classes. We also show some failed examples of our method in (g) and (h). Different colored boxes denote different classes, *i.e.* red box denotes 'Car', blue box denotes 'Pedestrian', cyan box denotes 'Bus', yellow box denotes 'Traffic Sign' and green box denotes 'Traffic Light'. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Computational complexity of different methods. Latency is measured as the average inference time per image (mean ± std) over 200 runs after 50 warm-up iterations. Throughput denotes one image (batch size = 1) processed per second under the same setting. All experiments are conducted on a single RTX3090 GPU.

Method	Parameters (M)	Latency ↓ (ms)	Throughput ↑ (images/s)	FLOPs (G)	Memory (MB)
2PCNet [11]	165.22	34.64 ± 0.49	28.88	73.61	1192.7
ISP-Teacher [15]	173.62	33.40 ± 0.35	29.94	73.61	1226.7
Cos [20] (SOTA)	269.07	34.74 ± 0.42	28.79	73.61	1589.2
Ours	277.48	33.73 ± 0.41	29.65	73.61	1623.6

ing based ISP degradation loss L_{self} into the original detection loss L_{det}. To explore the influence of the weights β of L_{self}, we set different values of $\beta = 1, 2, 5$ to conduct experiments on the BDD100k dataset. As shown in Table 4, when $\beta = 1$, we get the best performance not only in AP (49.6%) but also in small-sized object (9.3% in AP_S), medium-sized object (27.3% in AP_M) and large-sized object (46.2% in AP_l). However, the performance declines slightly when $\beta = 2$ (AP from 49.6% to 49.2%), and there is a significant decrease in AP performance when $\beta = 5$, *i.e.* 46.8%, which is even lower than the original version method ISP-Teacher [15] (48.8% in AP). The above experiments indicate that the weight of self-supervised learning based ISP degradation loss is sensitive to the performance of object detection, and we set $\beta = 1$ by default in this paper.

iii) Analysis of mixup ratio ρ in Mixup augmentation (Eq. (5)). As shown in Fig. 7, the detection performance on both BDD100k and SHIFT datasets exhibits a consistent trend with respect to the mixup ratio (ρ). When ρ increases from 0.2 to 0.5, the AP values gradually improve, reaching the best performance at $\rho = 0.5$ (49.6% AP on BDD100k and 52.8% AP on SHIFT). This indicates that a moderate mixing ratio effectively enhances representation diversity between base and mixed instances. However, further increasing ρ beyond 0.5 leads to a slight performance decline. Therefore, we set $\rho = 0.5$ as default, which provides a balanced contribution between the base instance and the mixed instance in the CRM.

4.5. Computational complexity analysis

From Table 5 we can see that the latency of proposed ISP Dynamic Teacher is 33.73 ms per image, which is similar to ISP-Teacher [15] (33.40 ms) and slightly lower than Cos (34.74 ms). All methods share the same FLOPs (73.61 G), as they use the same Faster R-CNN [5] detector during inference. This confirms that the additional complexity is mostly in the training phase (*e.g.* teacher-student learning or auxiliary modules), and does not affect the forward pass computational cost at test time. Compared with SOTA method Cos [20], our method achieves higher throughput (29.65 *vs.* 28.79 images/s) and lower inference latency (33.73 ms *vs.* 34.74 ms) with similar parameters and GPU memory usage.

4.6. Visualization results on BDD100k dataset

To show the effectiveness of our ISP Dynamic Teacher, we present some visualization results on BDD100k val dataset. As shown in Fig. 8,

the first two rows are dark object detection results in clear weather. From (a) and (b) in Fig. 8, we see that our ISP Dynamic Teacher can detect all objects accurately. However, AT [10] and 2PCNet [11] sometimes detect wrong cars, *i.e.* extra red box, while ISP-Teacher [15] and Cos [20] mistakenly detect some cars and traffic light (green box), respectively. Compared with the first two rows, (c) is more challenging because it is in a rainy scene. Under such adverse weather conditions, our method also achieves better results. However, AT [10] and Cos [20] miss the pedestrian on the left of the picture while ISP-Teacher [15] even misses all the pedestrians and cars on the left of the picture.

Furthermore, the visualization results in (d), (e) and (f) in Fig. 8 are the best proof of our newly added contributions, *i.e.* class-rebalanced module and dynamic IoU-threshold adjustment. From Fig. 2, we can see that 'Bus' is a minority class and there is a huge gap between the number of 'Bus' and other majority classes (*e.g.* 'Car', 'Traffic Sign' and 'Traffic Light'). This serious class imbalance could affect the object detection results, as shown in (d) of Fig. 8, AT [10] cannot detect the existence of 'Bus' class at all. Although 2PCNet [11], ISP-Teacher [15] and Cos [20] can locate part of objects, these methods misclassify the 'Bus' class as 'Car' class (small red box in the middle of figure). However, our method can locate and detect the 'Bus' class (two cyan boxes in picture) thanks to the proposed class-rebalanced module. Lines (e) and (f) of Fig. 8 are the detection results for small instances. From (e), we can see that all methods can detect 'Traffic Sign' (yellow boxes), however, compared with 'Car' that are smaller than 'Traffic Sign', other methods have a certain degree of missed detection while our method could detect all cars. Line (f) is more challenging, as 'Traffic Sign' and 'Traffic Light' are almost invisible to the human eyes in dark condition. From the previous analysis, we know that a higher static threshold may filter out very small instances, which makes it more difficult for the model to learn representations of these objects. By elaborately designing dynamic IoU-threshold adjustment strategy, our method does not miss any small instances in dark conditions.

5. Conclusion and future work

In this paper, we propose a novel dark object detection method, named ISP Dynamic Teacher, for challenging low-light scenes without annotations. To overcome the problem that mainstream teacher-student architecture based UDA methods have poor results on the day-to-night condition, we design a day-to-night transformation module that is consistent with the ISP pipeline of the camera sensors (ISP-DTM) to make the augmented images look more in line with the natural low-light images captured by the cameras. Moreover, a self-supervised learning strategy is used to capture the intrinsic visual information of images under different light changes. In order to avoid self-supervised learning based ISP degradation affecting the training process of object detection, a disentanglement regularization is introduced by minimizing the cosine similarity of the gradients of different tasks while maximizing the gradients of the same tasks.

Furthermore, based on the above contributions, we propose a Class-rebalanced Module (CRM) by creating a confusion matrix and utilizing a weight sampling algorithm to find the mix instance that best matches the base instance in the original input image. And then we mix them together using Mixup augmentation, which can solve the problem of class imbalance in dark object detection. Building upon this, we further devise a Dynamic IoU-threshold Adjustment (DIA) strategy to find the True-Positive pseudo-labels that are filtered out mistakenly by a high static threshold. Experimental results on two benchmarks show that our method outperforms previous teacher-student architecture methods in dark scenes by a large margin.

However, our method also has some limitations. i) similar to other pseudo-labeling strategies, the proposed DIA method has inherent limitations. In particular, DIA relies on teacher-student IoU consistency as a proxy for localization reliability, and thus may be susceptible to false positives in cases where student and teacher detections align in IoU but belong to different classes. We believe that incorporating additional semantic consistency cues (*e.g.* feature similarity constraints) is a promising direction for future work. In addition, ii) the performance may be sensitive to ISP parameter settings and domain-specific characteristics such as illumination distribution, sensor response, or noise patterns. Future work will explore domain-aware parameter fine-tuning to enhance the robustness of the ISP module across diverse conditions.

CRediT authorship contribution statement

Yin Zhang : Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization; **Yongqiang Zhang :** Writing – review & editing, Methodology, Investigation, Funding acquisition; **Zian Zhang :** Validation, Methodology, Data curation; **Man Zhang :** Validation, Investigation; **Rui Tian :** Visualization, Software, Methodology; **Mingli Ding :** Investigation, Funding acquisition; **Bogdan Raducanu :** Writing – review & editing, Supervision; **Dan Liu :** Supervision, Funding acquisition.

Data availability

The data used in this study are publicly available.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the Inner Mongolia Natural Science Foundation for Distinguished Young Scholars (No. 2025JQ009), the Inner Mongolia Talent Development Project for Outstanding Young Talents. This work is supported by Grant PID2022-143257NB-I00 funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU, the Departament de Recerca i Universitats from Generalitat de Catalunya with reference 2021SGR01499, and the Generalitat de Catalunya CERCA Program.

References

[1] J. Zhan, Y. Luo, C. Guo, Y. Wu, J. Meng, J. Liu, YOLOPX: anchor-free multi-task learning network for panoptic driving perception, Pattern Recognit. 148 (2024) 110152.
[2] F. Zhong, Y. Wu, H. Yu, G. Wang, Z. Lu, A benchmark dataset and semantics-guided detection network for spatial–temporal human actions in urban driving scenes, Pattern Recognit. 158 (2025) 111035.
[3] X. Wang, X. Yang, S. Zhang, Y. Li, L. Feng, S. Fang, C. Lyu, K. Chen, W. Zhang, Consistent-teacher: towards reducing inconsistent pseudo-targets in semi-supervised object detection, The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) (2023).
[4] M. He, Y. Wang, J. Wu, Y. Wang, H. Li, B. Li, W. Gan, W. Wu, Y. Qiao, Cross domain object detection by target-perceived dual branch distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9570–9580.
[5] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).
[6] Z. Cui, Y. Zhu, L. Gu, G.-J. Qi, X. Li, R. Zhang, Z. Zhang, T. Harada, Exploring resolution and degradation clues as self-supervised signal for low quality object detection, in: European Conference on Computer Vision, Springer, 2022, pp. 473–491.
[7] Z. Cui, G.-J. Qi, L. Gu, S. You, Z. Zhang, T. Harada, Multitask AET with orthogonal tangent regularity for dark object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2553–2562.
[8] P. Mi, J. Lin, Y. Zhou, Y. Shen, G. Luo, X. Sun, L. Cao, R. Fu, Q. Xu, R. Ji, Active teacher for semi-supervised object detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[9] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, P. Vajda, Unbiased teacher for semi-supervised object detection, in: Proceedings of the International Conference on Learning Representations (ICLR), 2021.
[10] Y.-J. Li, X. Dai, C.-Y. Ma, Y.-C. Liu, K. Chen, B. Wu, Z. He, K. Kitani, P. Vajda, Cross-domain adaptive teacher for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[11] M. Kennerley, J.-G. Wang, B. Veeravalli, R.T. Tan, 2PCNet: two-phase consistency training for day-to-night unsupervised domain adaptive object detection, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
[12] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, T. Darrell, et al., Bdd100k: a diverse driving video database with scalable annotation tooling, 2 (5) (2018) 6 arXiv:1805.04687.
[13] T. Sun, M. Segu, J. Postels, Y. Wang, L. Van Gool, B. Schiele, F. Tombari, F. Yu, SHIFT: a synthetic driving dataset for continuous multi-task domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21371–21382.
[14] H. Jang, K. Sohn, Enhancing source-free domain adaptive object detection with low-confidence pseudo label distillation, in: European Conference on Computer Vision, 2024.
[15] Y. Zhang, Y. Zhang, Z. Zhang, M. Zhang, R. Tian, M. Ding, ISP-Teacher: image signal process with disentanglement regularization for unsupervised domain adaptive dark object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 7387–7395.
[16] G. Sheng, G. Hu, X. Wang, W. Chen, J. Jiang, Low-light image enhancement via clustering contrastive learning for visual recognition, Pattern Recognit. 164 (2025) 111554.
[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-End Object Detection with Transformers, Springer, 2020.
[18] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, L. Zhang, Image-adaptive YOLO for object detection in adverse weather conditions, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 1792–1800.
[19] J. Lu, W. Zheng, Y. Qian, X. Liang, K. Wang, Z. Yuan, Illumination-adaptive feature enhancement for low-light object detection, Pattern Recognit. 176 (2026) 113122.
[20] J. Yuan, A. Le-Tuan, M. Hauswirth, D. Le-Phuoc, Cooperative students: navigating unsupervised domain adaptation in nighttime object detection, (2024) arXiv:2404.01988.
[21] Y. Cui, L. Li, H. Yin, Y. Gao, Y. Sun, C. Yan, Debiased teacher for day-to-night domain adaptive object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 2577–2587.
[22] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, CutMix: regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
[23] S.-A. Choe, A.-H. Shin, K.-H. Park, J. Choi, G.-M. Park, Open-set domain adaptation for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23943–23953.
[24] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, C. Finn, Gradient surgery for multi-task learning, Adv. Neural Inf. Process. Syst. 33 (2020) 5824–5836.
[25] M. Suteu, Y. Guo, Regularizing deep multi-task networks using orthogonal gradients, (2019) arXiv:1912.06844.
[26] E. Zhang, C. Geng, S. Chen, Class-aware universum inspired re-balance learning for long-tailed recognition, Pattern Recognit. 161 (2025) 111337.
[27] F. Zhang, T. Pan, B. Wang, Semi-supervised object detection with adaptive class-rebalancing self-training, in: Proceedings of the AAAI Conference on Artificial Intelligence, 3, 2022, pp. 3252–3261.
[28] M. Kennerley, J.-G. Wang, B. Veeravalli, R.T. Tan, CAT: Exploiting inter-class dynamics for domain adaptive object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16541–16550.
[29] P.E. Debevec, J. Malik, Recovering high dynamic range radiance maps from photographs, in: Seminal Graphics Papers: Pushing the Boundaries, Vol. 2, 2023, pp. 643–652.
[30] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, J.T. Barron, Unprocessing images for learned raw denoising, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11036–11045.
[31] C. Sakaridis, D. Dai, L. 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, 2021, pp. 10765–10775.
[32] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[33] C. Sakaridis, D. Dai, L. Van Gool, Semantic foggy scene understanding with synthetic data, Int. J. Comput. Vis. 126 (9) (2018) 973–992.
[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[35] J. Deng, W. Li, Y. Chen, L. Duan, Unbiased mean teacher for cross-domain object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4091–4101.

Yin Zhang received his M.S. degree in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2021. He is currently a Ph.D. student from the Harbin Institute of Technology (HIT) and a visiting student in Computer Vision Center (CVC) from Universitat Autònoma de Barcelona (UAB). His research areas are computer vision, pattern recognition, machine learning, and deep learning. His research interests mainly include object detection, domain adaptation, knowledge distillation.

Yongqiang Zhang received M.S. and Ph.D. degrees in instrument science and technology from the Harbin Institute of Technology (HIT), Harbin, China, in 2015 and 2020, respectively. He worked at King Abdullah University of Science and Technology (KAUST) as a visiting student from 2017 to 2018. He was an assistant professor in the School of Instrumentation Science and Engineering at the Harbin Institute of Technology from 2020 to 2024. Currently, he is a professor in the College of Computer Science (College of Software-College of Artificial Intelligence) at Inner Mongolia University. His research areas are computer vision, pattern recognition, machine learning, and deep learning. His research interests mainly include face detection, weakly/fully supervised object detection, activity detection, and image and video understanding in the real world.

Zian Zhang received his B.S. and M.S. degrees in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2018 and 2020, respectively. He worked at King Abdullah University of Science and Technology (KAUST) as a visiting student in 2024. Currently, he is a Ph.D. student in the School of Instrumentation Science and Engineering at HIT. His research areas are computer vision, pattern recognition, machine learning, and deep learning. His research interests mainly include human pose estimation, object detection, skeleton-based biometric recognition.

Man Zhang received his M.S. degree in Optical Engineering from Harbin Institute of Technology (HIT), Harbin, China, in 2020. He is currently a Ph.D. student from the Harbin Institute of Technology (HIT). His research areas are computer vision, pattern recognition, machine learning, and deep learning. His research interests mainly include object detection, motion forecasting and 3D occupancy prediction.

Rui Tian received the MS degree in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 2023. He is currently a PhD student from the Harbin Institute of Technology (HIT). His research areas are computer vision, pattern recognition, machine learning, and deep learning. Author Biography His research interests mainly include weakly supervised object detection, depth estimation and contrastive learning.

Mingli Ding received the B.S., M.S. and Ph.D. degrees in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 1996, 1997 and 2001, respectively. He worked as a visiting scholar in France from 2009 to 2010. Currently, he is a professor in the School of Instrumentation Science and Engineering at Harbin Institute of Technology. Prof. Ding's research interests are intelligence tests and information processing, automation test technology, computer vision, and machine learning. He has published over 40 papers in peer-reviewed journals and conferences.

Bogdan Raducanu received the Ph.D. degree "Cum Laude" from the University of the Basque Country (UPV/EHU), Spain, in 2001. Between 2002-2004 he was a post-doc researcher at Technical University of Eindhoven (TU/e), The Netherlands. In 2004 he joined CVC as project director, after obtaining a 'Ramon y Cajal' fellowship. Currently, he is a senior researcher and project director at Computer Vision Center. His research interests are in the areas of deep learning and computer vision with applications to pattern recognition, social signal processing, robotics and wearable computing. He has published more than 100 papers in peer-reviewed international conferences and journals and participated in more than 20 regional, national and international projects.

Dan Liu received B.S, MS. and Ph.D. degrees in instrument science and technology from the Harbin Institute of Technology (HIT), Harbin, China, in 2002, 2004 and 2008, respectively. He worked as a visiting scholar in USA from 2014 to 2015. Currently, he is a associate professor in the School of Instrumentation Science and Engineering at the Harbin Institute of Technology. Prof. Liu's research interests are weak signal detection, intelligence tests and data processing, multiple-functional sensing, and machine learning. He has published over 80 papers in peer-reviewed journals and conferences.