ISP-Teacher:Image Signal Process with Disentanglement Regularization for Unsupervised Domain Adaptive Dark Object Detection

Yin Zhang[*], Yongqiang Zhang[*], Zian Zhang, Man Zhang, Rui Tian, Mingli Ding

School of Instrument Science and Engineering, Harbin Institute of Technology
{yin.zhang.hit, yongqiang.zhang.hit, sieann.hit, man.zhang.hit, rui.tian.hit, mingli.ding.hit}@gmail.com

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 neglected to model the intrinsic imaging process, *i.e.* image signal processing (ISP), which is important for camera sensors to generate low-light images. To solve the above problems, in this paper, we propose a novel method named ISP-Teacher for dark object detection by exploring Teacher-Student architecture from a new perspective (*i.e.* self-supervised learning based ISP degradation). Specifically, we first design a day-to-night transformation module that 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. Moreover, 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 gradients vectors as orthogonal as possible. Extensive experiments conducted on two benchmarks show the effectiveness of our method in dark object detection. In particular, ISP-Teacher achieves an improvement of +2.4% AP and +3.3% AP over the SOTA method on BDD100k and SHIFT datasets, respectively. The code can be found at https://github.com/zhangyin1996/ISP-Teacher.

Introduction

Object detection has achieved remarkable success and widely used in various fields such as security monitoring and autonomous driving. However, these object detection models trained on high-quality daytime images often perform poorly on low-light images, because these images taken under dark conditions suffer from various types of light and undesirable noise (Cui et al. 2022a). Furthermore, annotating low-light images is also difficult, so it is impossible to obtain high-quality annotation information of low-light images like daytime images. At present, dark object detection is still an urgent problem to be solved.

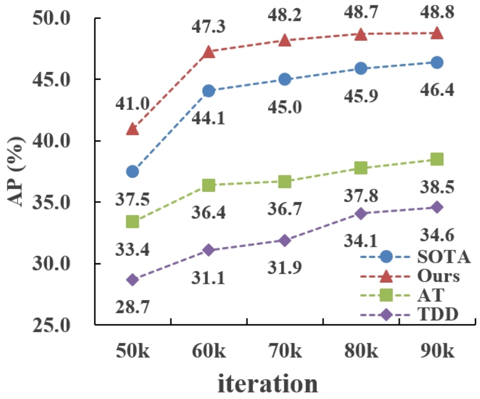

Figure 1: Results of some Teacher-Student architecture UDA methods on BDD100k. We found that AT and TDD get worse results on day-to-night conditions, which even lower than the baseline detector Faster-RCNN (41.1% AP). Our proposed method outperforms other counterparts by a large margin and always higher than SOTA method (Kennerley et al. 2023) in any iteration.

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, the enhanced images are visually comfortable for humans but do not benefit to the high-level task for machine vision (Cui et al. 2022b, 2021). To this end, unsupervised domain adaptive (UDA) has been proposed to address this problem.

Recently, Teacher-Student architecture (Sohn et al. 2020) has attracted lots of attention in semi-supervised object detection (Wang et al. 2023; Mi et al. 2022; Liu et al. 2021) and has also achieved excellent results in the field of domain adaptation object detection (Li et al. 2022; He et al. 2022; Kennerley et al. 2023). However, as shown in Figure 1, we found that the best Teacher-Student UDA methods like AT (Li et al. 2022), TDD (He et al. 2022) achieve good results on regular domain adaptive datasets (*e.g.* Cityscapes to Foggy Cityscapes) but get poor results on day-to-night conditions. They are even lower than the baseline detector

[*] These authors contributed equally.

(*i.e.* Faster-RCNN) that trained on daytime images and directly applied to nighttime images (41.1% AP).

We think there are some special difficulties for object detection in dark conditions: i) Low-light images from camera sensors suffer from imbalanced noise, exposure, lighting and blur *etc.* , which are not found in daytime images, thus training the student network with these daytime images inevitably produce some domain bias. ii) Most of these Teacher-Student based methods solve the domain bias problem by optimizing the framework (He et al. 2022), selecting useful ground-truth labels information (Mi et al. 2022) or revising the score threshold of pseudo-bboxes dynamically (Wang et al. 2023). They usually imitate the dark conditions by using traditional non-learnable data augmentation strategies on the available annotated source daytime images. However, these methods neglected to model the intrinsic imaging process, *i.e.* image signal processing (ISP), which is important for camera sensors to generate low-light images.

In this paper, we solve the above problems by exploring 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 the intrinsic visual information that is not affected by lighting changes, which can address the domain bias of student network.

First, inspired by image signal process (ISP) pipeline, which is a crucial component in cameras that transforms RAW data into RGB images for person visualization (Yu et al. 2021; Cui et al. 2021). We replace traditional non-learnable data augmentation with self-supervised learning based ISP degradation, where a day-to-night transformation module that consistent with the ISP pipeline of the camera sensors (ISP-DTM) is proposed to obtain the low-light images from daytime images. Then, an Encoder-Decoder structure is utilized to encode the pair of daytime and nighttime images and decodes them into some parameters, such as gamma, light intensity, *etc.* through a self-supervised learning manner. Thus, the intrinsic visual information can be learned under the supervision of L_1 loss.

However, joint training of the self-supervised learning based ISP degradation and object detection task in a shared encoder may cause over-entanglement problem (*i.e.* gradient conflict problem). We found these two tasks have a negative cosine similarity that will hurt the final performance. To this end, we propose a disentanglement regularization by minimizing the gradients of cosine similarity of self-supervised learning based ISP degradation and object detection while maximizing cosine similarity of the same tasks. This simply implement can push two gradients vectors as orthogonal as possible and make the two tasks not affect each other.

To sum up, the contributions of this paper are as follows:

- A novel dark object detection method named ISP-Teacher is proposed from a new perspective to explore a self-supervised learning based ISP degradation in a Teacher-Student architecture, which could adapt to challenging low-light conditions in the real world.
- We design a day-to-night transformation (ISP-DTM) module 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 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 simply implement could decouple these two tasks in a shared encoder.
- Extensive experiments conducted on BDD100k and SHIFT datasets show the effectiveness of our proposed method. In particular, ISP-Teacher achieves the new best performance on BDD100k and SHIFT datasets by improving +2.4% and +3.3% in AP over the state-of-the-art method, respectively.

Related Work

Object Detection in Dark Conditions

To tackle the problem of object detection in low-light conditions, a direct way is use low-light enhancement methods (Guo et al. 2020; Jin et al. 2023; Wu et al. 2023) to process the dark images and then send the de-dimming images to the mainstream object detection methods (Ren et al. 2015; Redmon and Farhadi 2018; Carion et al. 2020) for inference. However, the detection performance of these methods is unsatisfactory on some natural dark images. As a result, some end-to-end methods that train the low-light enhancement and object detection tasks jointly. For example, IA-YOLO (Liu et al. 2022) 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 (Cui et al. 2021) introduces a multitask auto encoding transformation model to decode low-light degrade transformation by considering noise and ISP pipeline in cameras. The main difference between MAET and our work is that we regard the ISP degradation as a self-supervised learning task for Teacher-Student domain adaptive object detection. Furthermore, although 2PCNet (Kennerley et al. 2023) is a nighttime domain adaptive object detection method, it proposes a non-learnable data augmentation while our method is self-supervised and we consider the principle of the camera sensor in ISP-DTM.

Disentanglement Regularization

Multi-task networks usually contain an encoder and several decoders for specific tasks. However, these approaches face an optimization problem which sometimes leading 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, (Yu et al. 2020) 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. (Suteu and Guo 2019) finds nearly orthogonal gradients would not interfering with each other tasks, and proposes that regularizing the angle between gradients to solve the negative trans-

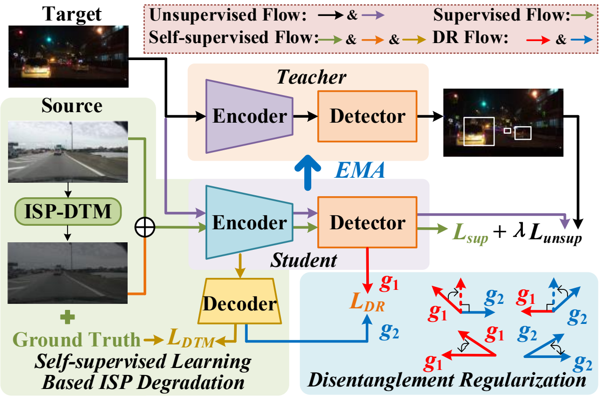

Figure 2: The architecture of our ISP-Teacher. Our pipeline is based on the Teacher-Student architecture, the green area (left side) is the proposed self-supervised learning based ISP degradation and the blue area (right side) is the illustration of disentanglement regularization.

fer problem. The above methods are focus on classification and regression tasks, and our work is inspired by recent research (Cui et al. 2021) that minimizes the absolute value of cosine similarity to disentangle the object detection and degrade transformation tasks. Different from (Cui et al. 2021), 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.

Proposed Method

Overview of ISP-Teacher

Let $D_{day} = \{X_l, Y_l\}$ denotes the daytime dataset, which contains X_l daytime images with Y_l labels in source domain. $D_{night} = \{X_u\}$ denotes the nighttime dataset, and it only contains X_u nighttime images without labels in target domain. Subscript l and u indicate the labeled and unlabeled data, respectively. As shown in Figure 2, our ISP-Teacher consists of a student network and a teacher network. Similar to prior works (Kennerley et al. 2023; Liu et al. 2021), both student and teacher are the Faster-RCNN (Ren et al. 2015) structure, and the detection loss L_{det} is as following:

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

where L_{sup} and L_{unsup} denote the supervised learning loss and unsupervised learning loss, respectively.

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 (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. After the burn up stage, all the weights of student are transferred to the teacher.

The teacher network only takes nighttime images (target domain) as 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 Y_p denotes pseudo-labels. Noted that the unsupervised loss is only applied in the classification while not used in the bounding box regression.

Furthermore, there are two components in the proposed ISP-Teacher. The first component is the self-supervised learning based ISP degradation (green area on the left of Figure 2), which is used to capture the intrinsic visual information that is not affected by lighting changes. The second component is the disentanglement regularization (DR, blue area on the right of Figure 2), which disentangles dark object detection and self-supervised learning based ISP degradation to mitigate the impact between each other. In the next subsection, we will illustrate our proposed self-supervised learning based ISP degradation and DR in details.

Self-supervised Learning Based ISP Degradation

Self-supervised learning based ISP degradation contains a day-to-night transformation module that consistent with the ISP pipeline of the camera sensors (ISP-DTM) and a self-supervised learning strategy.

ISP-DTM As shown in Figure 3, 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. Base 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 are restored back to sRGB by the ISP Pipeline step for dark object detection. As shown in green area of Figure 3, 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 details:

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 appearing to be lit under the neutral illumination (Cui et al. 2021). The

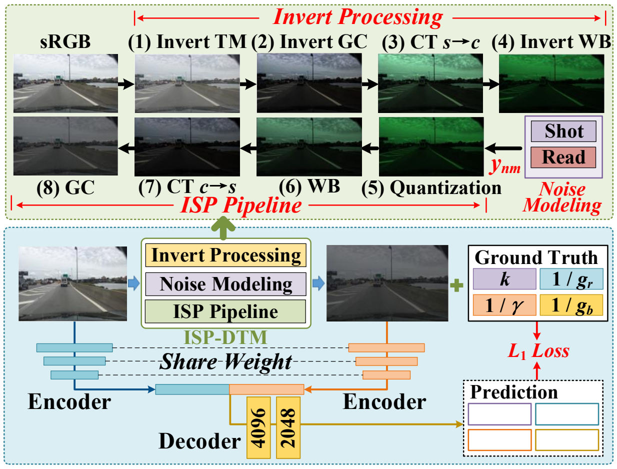

Figure 3: The structure of self-supervised learning based ISP degradation.

detailed process is as follows:

$$\begin{bmatrix} \hat{I}_r \\ \hat{I}_g \\ \hat{I}_b \end{bmatrix} = \begin{bmatrix} g_r & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & g_b \end{bmatrix} \cdot \begin{bmatrix} I_r \\ I_g \\ I_b \end{bmatrix} \tag{4}$$

where I and $\hat{I}$ denote the image before and after white balance respectively, and the subscripts r, g, b represent the three channels of the RGB image. The channel gain of red g_r and blue g_b are random sampled from (1.9, 2.4) and (1.5, 1.9) uniformly and independently, and set $1/g_r$ and $1/g_b$ in invert process based on (Brooks et al. 2019).

Color Transformation. Because the data format of standard color space (sRGB) do not match the camera internal color space (cRGB), we use a 3 × 3 color correction matrix T_{ccm} to achieve this color transformation:

$$y_{c\to s} = T_{ccm} \cdot I_{cRGB} \tag{5}$$

$$y_{s\to c} = T_{ccm}^{-1} \cdot I_{sRGB} \tag{6}$$

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, where the dark areas of the pixels have a larger change rate and the light areas of the pixels have a smaller change rate. 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:

$$I_{out} = \alpha I_{in}^{1/\gamma} \tag{7}$$

where I_{in} and I_{out} denotes input and output images, α and γ are used to adjust the shape of gamma correction curve. When γ is less than 1, the overall image will be stretched in the direction of strong illumination, and when γ value is greater than 1, it will be stretched in the direction of weak illumination. In this paper, γ is sampled from an uniform distribution $\gamma \sim U(2, 3.5)$ and α is set to 1. The invert process of gamma correction is to replace $1/\gamma$ in Eq.7 with γ:

$$I_{out} = \alpha I_{in}^{\gamma} \tag{8}$$

Tone Mapping. High dynamic range images in real scenes require tone mapping operation to suit the dynamic range of camera sensors (Debevec and Malik 2023). Usually, the tone mapping process includes three steps: first calculating the average brightness of current scenes, then selecting a suitable brightness area according to the average brightness, and finally mapping the entire scene to this brightness area to get a correct result. Here, we simplify the tone mapping to a simple 'smoothstep' curve:

$$F_{tm}(x) = 3x^2 - 2x^3 \tag{9}$$

and it invert process is:

$$F_{tm}^{-1}(y) = \frac{1}{2} - \sin(\frac{\sin^{-1}(1-2y)}{3}) \tag{10}$$

Noise Modeling. Noises of camera sensors primarily comes 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 the cameras. Mathematically, shot noise is a Poisson random variable whose mean is the light intensity (*i.e.* parameter k in Eq.11 and Eq.14) and read noise is a Gaussian random variable with zero mean and fixed variance (Brooks et al. 2019). We model both of them as x_{noise}:

$$x_{noise} \sim N(\mu = 0, \sigma^2 = k \cdot I \cdot \lambda_{shot} + \lambda_{read}) \tag{11}$$

where I is the output image from 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 (Brooks et al. 2019). The details of the sampling process are as follows:

$$\log \lambda_{shot} \sim U(a = \log(0.0001), b = \log(0.012)) \tag{12}$$

$$\log \lambda_{read} \sim N(\mu = 2.18 \log \lambda_{shot}, \sigma = 0.26) \tag{13}$$

Moreover, parameter k in Eq.11 is the light intensity (between 0.01 and 1.0), and it follows a truncated Gaussian distribution (Cui et al. 2021):

$$k \sim N(\mu = 0.1, \sigma = 0.08) \tag{14}$$

Finally, the outputs y_{nm} of Noise Modeling step can be formulated as:

$$y_{nm} = k \cdot I + x_{noise} \tag{15}$$

which is then sent to ISP Pipeline step for subsequent transformation processing.

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

$$\hat{y} \sim (-\frac{1}{2B}, \frac{1}{2B}) \tag{16}$$

$$y_{quant} = y_{nm} + \hat{y} \tag{17}$$

where B is randomly selected from 12, 14 and 16 as in (Cui et al. 2021).

Moreover, during the process of ISP-DTM, we calculate four parameters, *i.e.* light intensity k in Eq.14, $1/gamma$ in (2) Invert Gamma Correction, channel gain $1/g_r$ and $1/g_b$ in (4) Invert White Balance, which are used as the ground truth in the following self-supervised learning strategy.

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

$$I_l + p_i = T_{ISP} \left[T_{invert}(I) + y_{nm}\right] \tag{18}$$

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 image into high-level features by a weight-shared Encoder, and then to decode four parameters $\tilde{p}_i (i = 1, 2, 3, 4)$ as the degradation predictions. The loss of self-supervised learning strategy L_{self} is a L_1 loss:

$$L_{self} = \frac{1}{4} \sum_{i=1}^{4} L_1(p_i, \tilde{p}_i) \tag{19}$$

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 weight of k, $1/gamma$, $1/g_r$, $1/g_b$ are set to 5:1:1:1 in our implementation.

Disentanglement Regularization (DR)

As shown in Figure 2, 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 gradient.

To overcome this issue, we propose a regularization to disentangle these two tasks (*i.e.* ISP degradation and object detection) in the training process. The goal of our disentanglement regularization is that the gradients g_1 and g_2 of two different tasks have the minimum cosine similarity, *i.e.* the angle between two vectors tends to 90 degrees and the value of cosine closes to 0, while the gradient of the same tasks has the maximize cosine similarity.

Specifically, as shown in the bottom part of Figure 2, the red arrow g_1 and blue arrow g_2 denote gradients vectors of the task of object detection and self-supervised learning based ISP degradation respectively. For different tasks, we minimum cosine similarity by pushing the g_1 or g_2 close to dotted line arrow under the supervision of DR loss L_{DR}. For the same tasks, we make the gradients vectors as coincident as possible. Mathematically, DR can be expressed as:

$$\begin{aligned} L_{DR} =& \omega_1 \left|\cos(g_1, g_2)\right| + \omega_2(|1 - \cos(g_1, g_1)|) + \\ & \omega_3(|1 - \cos(g_2, g_2)|) \end{aligned} \tag{20}$$

where ω_1, ω_2 and ω_3 are the parameters to balance these three terms. In this paper, we set $\omega_1 = 5$ and $\omega_2 = \omega_3 = 0.5$.

Total Loss

The total loss function includes L_{self} loss that makes the encoder to capture the intrinsic visual information, L_{DR} pushes two gradients vectors at different tasks as orthogonal as possible, and L_{sup} and L_{unsup} are the original detection losses L_{det} in the Teacher-Student architecture, which can be formulated as:

$$L_{total} = \beta L_{self} + L_{DR} + L_{det} \tag{21}$$

where β is the weight of self-supervised learning loss in Eq.19.

Experiments

Datasets and Metrics

BDD100k (The Berkeley Deep Drive 100k) is a widely used autonomous driving dataset (Yu et al. 2018), which consists 70k training images, 20k test images and 10k validation images. It includes 10 common classes and covers various weather scenarios, *e.g.* rainy, snowy, foggy, overcast and *etc.* . Following (Kennerley et al. 2023), we split 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 validation dataset are used for validation. After splitting, there are 36728 daytime images and 32998 nighttime images in the training set and 4707 nighttime images in the validation set.

SHIFT is also an autonomous driving dataset (Sun et al. 2022), 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 has the same 6 classes as BDD100k with bounding box annotations. similar to BDD100k, we also split it into 19452 daytime images and 8497 nighttime images for training and 1200 nighttime images for validation.

As for the metrics, following the method of (Kennerley et al. 2023), we adopt AP (*i.e.*AP$_{50}$, IoU@0.5), AP$_S$ (small-sized object), AP$_M$ (medium-sized object) and AP$_L$ (large-sized object) to evaluate our model.

Implementation Details

Following previous Teacher-Student architecture based domain adaption methods, we use Faster-RCNN (Ren et al. 2015) with ResNet50 (He et al. 2016) 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 hyperparameter λ = 0.3 and $\beta = 1$, and the rate smooth coefficient parameter of EMA is set to 0.9996. The batch size is 4, which includes 2 daytime images in source domain and 2 nighttime images in target, and all images are proportionally scaled to a minimum side of 600. For the burn up stage, we train the student network under a supervised manner on source domain for 50k and 20k iterations for BDD100k and SHIFT datasets, respectively. And the total iterations on BDD100k and SHIFT is 90k and 70k iterations. Our method is implemented based on *detectron2* (Wu et al. 2019) with 4 RTX6000 GPUs.

Method	AP	Ped.	Rid.	Car	Tru.	Bus	Mot.	Bic.	T-Light	T-Sign
Source (Lower-Bound)	41.1	50.0	28.9	66.6	47.8	47.5	32.8	39.5	41.0	56.5
Oracle (Upper-Bound)	46.2	52.1	35.0	73.6	53.5	54.8	36.0	41.8	52.2	63.3
UMT (Deng et al. 2021)	36.2	46.5	26.1	46.8	44.0	46.3	28.2	40.2	31.6	52.7
TDD (He et al. 2022)	34.6	43.1	20.7	68.4	33.3	35.6	16.5	25.9	43.1	59.5
AT (Li et al. 2022)	38.5	42.3	30.4	60.8	48.9	52.1	34.5	42.7	29.1	43.9
2PCNet (Kennerley et al. 2023)	46.4	54.4	30.8	**73.1**	53.8	55.2	37.5	44.5	49.4	65.2
ISP-Teacher (Ours)	**48.8**	**57.8**	**39.4**	72.9	**54.6**	**55.9**	**43.8**	**48.1**	**49.6**	**66.3**

Table 1: Main results of our proposed method on BDD100k dataset. We show the average precision (AP) of each class. The full classes name from left to right are Pedestrian, Rider, Car, Trunk, Bus, Motorcycle, Bicycle, Traffic Light and Traffic Sign.

Main Results

In order to verify the effectiveness of our ISP-Teacher for dark object detection, we compare our method with some SOTA methods, *i.e.* UMT (Deng et al. 2021), TDD (He et al. 2022), AT (Li et al. 2022) and 2PCNet (Kennerley et al. 2023). It should be emphasized that 2PCNet (Kennerley et al. 2023) is an object detection method specially designed for low-light images and it achieves SOTA performance on BDD100k and SHIFT datasets. For the fairness of comparison, all of methods use ResNet50 (He et al. 2016) as the backbone, and the results in our experiment are shown in Table 1. In addition, we report the results that training Faster-RCNN with only daytime images and test on nighttime images denotes as ‘Source’ (Lower-Bound). On the other hand, we also show the results of training Faster-RCNN on nighttime images with ground-truth and test on nighttime images denotes as ‘Oracle’ (Upper-Bound).

Experiments on BDD100k. On BDD100k dataset, compared to other Teacher-Student architecture based domain adaptive methods for dark object detection, ISP-Teacher achieves a better performance owning to the proposed self-supervised learning based ISP degradation and disentanglement regularization strategy. As shown in Table 1, the previous Teacher-Student architecture methods achieve terrible results on night scenes and even much lower than the Lower-Bound. By elaborately designing self-supervised learning based ISP degradation and disentanglement regularization strategy, our ISP-Teacher outperforms all the Teacher-Student architecture methods by a large margin. 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 +7.7% AP improvement (*i.e.* 48.8% *vs.* 41.1%). Furthermore, our unsupervised approach even outperforms the supervised method ‘Oracle’ (the second row) that trains on nighttime images with annotations by +2.6% in terms of AP. Compared with 2PCNet (Kennerley et al. 2023), which is a SOTA night-specific algorithm for dark object detection, our method also obtains an impressive improvement in AP (from 46.4% to 48.8%, +2.4%) and brings the best results in eight out of nine categories , where ‘Car’ is also only 0.2% lower.

Experiments on SHIFT. To further verify the effectiveness of our proposed method, we conduct experiments on 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. ISP-Teacher achieves an improvement of +3.3% AP compared to SOTA method 2PCNet, and +5.4% AP for ‘Oracle’. Furthermore, our method outperforms 2PCNet on all categories.

Method	AP	Ped.	Car	Tru.	Bus	Mot.	Bic.
Lower-B	41.6	40.4	44.5	49.9	53.7	14.3	46.7
Upper-B	47.0	49.7	51.5	56.0	53.6	19.2	52.4
UMT	31.1	7.7	47.5	18.4	46.8	16.6	49.2
AT	38.9	25.8	33.0	54.7	49.5	20.7	52.3
2PCNet	49.1	51.4	54.6	54.8	56.6	23.9	54.2
Ours	**52.4**	**51.6**	**59.1**	**58.7**	**62.3**	**24.1**	**58.3**

Table 2: Main results of our proposed method on SHIFT dataset. Lower-B and Upper-B denote Lower-Bound and Upper-Bound, respectively.

Ablation Study

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

Effectiveness of Self-supervised Learning Based ISP Degradation. We compare our self-supervised learning based ISP degradation (the third row of Table 3) with other methods: i) traditional non-learnable data augmentations like randomly color jittering, gray scaling, Gaussian blurring (the first row of Table 3), and ii) nighttime specific augmentations NightAug in 2PCNet (Kennerley et al. 2023) (the second row in Table 3). As shown in Table 3, we can see that the method of traditional non-learnable data augmentation conducted on daytime images obtains poor performance (42.2% AP) in low-light conditions. 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 based ISP degradation not only addresses the domain bias of student network but also explore how to learn intrinsic visual information of dark images, which achieves a huge improve-

NightAug	Self.	DR	AP	AP$_S$	AP$_M$	AP$_L$
-	-	-	42.2	7.9	23.0	38.8
✓	-	-	45.8	8.6	25.7	42.2
-	✓	-	48.5	**9.2**	27.1	45.2
-	✓	✓	**48.8**	**9.2**	**27.2**	**45.7**

Table 3: Ablation study of each component in our ISP-Teacher on BDD100k dataset. 'NightAug' denotes a non-learnable nighttime specific augmentation in 2PCNet. 'Self.' and 'DR' denote the self-supervised learning based ISP degradation and disentanglement regularization.

β	AP	AP$_S$	AP$_M$	AP$_L$
1	**48.8**	**9.2**	**27.2**	**45.7**
2	48.4	8.9	26.7	45.7
5	44.9	7.9	22.9	39.1

Table 4: The influence of different weight β in the self-supervised learning based ISP degradation loss.

ment in AP (from 42.2% to 48.5%, +6.3%). The above experiments can prove the effectiveness of our proposed self-supervised learning based ISP degradation on low-light images object detection.

Effectiveness of DR. As shown in the last row of Table 3, when adding disentanglement regularization (DR) in our framework, the detection performance can further improve to 48.8% from 48.5%. This is thanks to the disentangle of the object detection and self-supervised learning based ISP degradation.

Analysis of the weights β of self-supervised learning based ISP degradation loss. From Eq.21, we add the self-supervised learning 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 BDD100k dataset. As shown in Table 4, we get the best performance of 48.8% in AP when $\beta = 1$. However, when $\beta = 2$, the performance declines slightly, and there is a significant decrease in AP performance when $\beta = 5$, *i.e.* 44.9% which is even lower than the non-learnable method NightAug (45.8% 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.

Visualization Results

Visualization of ISP-DTM Pipeline. As shown in the green area of Figure 3, we show an example of ISP-DTM pipeline on BDD100k dataset. First, sRGB daytime images are converted to cRGB images by the reversing process. Then, shot and read noises are added to cRGB images, and cRGB is transformed into sRGB through the ISP Pipeline step for dark object detection. The augmented images look more in line with the natural low-light images captured by cameras.

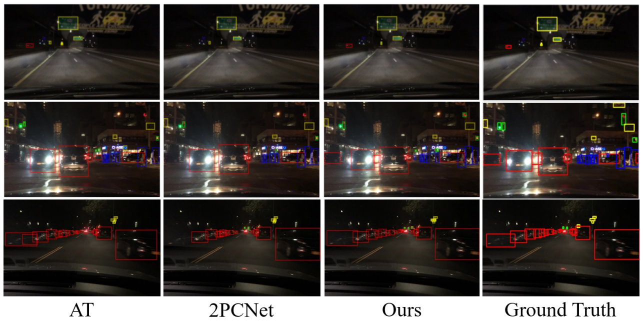

Figure 4: Examples of detection results on BDD100k dataset. From left to right: general Teacher-Student architecture UDA method AT (Li et al. 2022), SOTA method 2PCNet (Kennerley et al. 2023), our ISP-Teacher and Ground Truth. Different colored boxes denote different classes, *i.e.* red box denotes 'Car', blue box denotes 'Pedestrian', yellow box denotes 'Traffic Sign' and green box denotes 'Traffic Light'. Best seen on computer, in color and zoomed in.

Detection Results on BDD100k Dataset. Furthermore, to further show the effectiveness of our ISP-Teacher, we also present some visualization results on BDD100k val datasets. As shown in Figure 4, we can see that our ISP-Teacher could detect all objects accurately. However, AT (Li et al. 2022) mistakenly detects something as a traffic sign, *i.e.* an extra yellow box, and 2PCNet (Kennerley et al. 2023) misses a car (*i.e.* red box) in the first row. Moreover, as shown the second and third rows of Figure 4, our method also gets satisfactory results on complex scenes while other methods always have detection errors. For example, AT and 2PCNet miss some traffic light and cars in the second row.

Conclusion

In this paper, we propose a novel dark object detection method named ISP-Teacher for the 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 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 in our method by minimizing the cosine similarity of the gradients of different tasks while maximizing the gradients of the same tasks. Experimental results on two benchmarks show that our method outperforms previous Teacher-Student architecture methods in dark scenes by a large margin. However, object detection on low-light images is still a challenging task, *e.g.* the results on small object like traffic light need to be improved, and we plan to use Fourier-based mix strategy to learn more robust features for the student network in the future.

Acknowledgments

This work is supported by the China Postdoctoral Science Foundation (Grant No. 259822), the National Postdoctoral program for Innovative Talents (Grant No. BX20200108), and the National Science Foundation of China (Grant No. 62206077).

References

Brooks, T.; Mildenhall, B.; Xue, T.; Chen, J.; Sharlet, D.; and Barron, J. T. 2019. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11036–11045.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Cui, Z.; Li, K.; Gu, L.; Su, S.; Gao, P.; Jiang, Z.; Qiao, Y.; and Harada, T. 2022a. You Only Need 90K Parameters to Adapt Light: a Light Weight Transformer for Image Enhancement and Exposure Correction. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press.

Cui, Z.; Qi, G.-J.; Gu, L.; You, S.; Zhang, Z.; and Harada, T. 2021. Multitask AET With Orthogonal Tangent Regularity for Dark Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2553–2562.

Cui, Z.; Zhu, Y.; Gu, L.; Qi, G.-J.; Li, X.; Zhang, R.; Zhang, Z.; and Harada, T. 2022b. Exploring Resolution and Degradation Clues as Self-supervised Signal for Low Quality Object Detection. In *European Conference on Computer Vision*, 473–491. Springer.

Debevec, P. E.; and Malik, J. 2023. Recovering high dynamic range radiance maps from photographs. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 643–652.

Deng, J.; Li, W.; Chen, Y.; and Duan, L. 2021. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4091–4101.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1780–1789.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, M.; Wang, Y.; Wu, J.; Wang, Y.; Li, H.; Li, B.; Gan, W.; Wu, W.; and Qiao, Y. 2022. Cross domain object detection by target-perceived dual branch distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9570–9580.

Jin, X.; Han, L.-H.; Li, Z.; Guo, C.-L.; Chai, Z.; and Li, C. 2023. DNF: Decouple and Feedback Network for Seeing in the Dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18135–18144.

Kennerley, M.; Wang, J.-G.; Veeravalli, B.; and Tan, R. T. 2023. 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)*.

Li, Y.-J.; Dai, X.; Ma, C.-Y.; Liu, Y.-C.; Chen, K.; Wu, B.; He, Z.; Kitani, K.; and Vajda, P. 2022. Cross-Domain Adaptive Teacher for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; and Zhang, L. 2022. Image-adaptive YOLO for object detection in adverse weather conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1792–1800.

Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased Teacher for Semi-Supervised Object Detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mi, P.; Lin, J.; Zhou, Y.; Shen, Y.; Luo, G.; Sun, X.; Cao, L.; Fu, R.; Xu, Q.; and Ji, R. 2022. Active Teacher for Semi-Supervised Object Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*.

Sun, T.; Segu, M.; Postels, J.; Wang, Y.; Van Gool, L.; Schiele, B.; Tombari, F.; and Yu, F. 2022. SHIFT: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21371–21382.

Suteu, M.; and Guo, Y. 2019. Regularizing deep multi-task networks using orthogonal gradients. *arXiv preprint arXiv:1912.06844*.

Wang, X.; Yang, X.; Zhang, S.; Li, Y.; Feng, L.; Fang, S.; Lyu, C.; Chen, K.; and Zhang, W. 2023. Consistent-Teacher: Towards Reducing Inconsistent Pseudo-targets in Semi-supervised Object Detection. *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Wu, Y.; Pan, C.; Wang, G.; Yang, Y.; Wei, J.; Li, C.; and Shen, H. T. 2023. Learning Semantic-Aware Knowledge Guidance for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1662–1671.

Yu, F.; Xian, W.; Chen, Y.; Liu, F.; Liao, M.; Madhavan, V.; Darrell, T.; et al. 2018. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5): 6.

Yu, K.; Li, Z.; Peng, Y.; Loy, C. C.; and Gu, J. 2021. Reconfigisp: Reconfigurable camera image processing pipeline. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4248–4257.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33: 5824–5836.