TripleA: An Unsupervised Domain Adaptation Framework for Nighttime VRU Detection

Yuankun Wang, Zhenfeng Shao, Jiaming Wang, Yu Wang, Yulin Ding, Gui Cheng

***Abstract*—Detecting vulnerable road users (VRUs) at night presents significant challenges. Numerous methods rely heavily on annotations, yet the low visibility of nighttime images poses difficulties for labeling. To obviate the need for nighttime annotations, unsupervised domain adaptation manifests as a viable solution. However, existing approaches primarily focus on semantic-level domain gaps, often overlooking pixel-level discrepancies caused by inherent degradations in the nighttime domain. These degradations can impair machine vision and limit detection performance. In this paper, we propose TripleA, an unsupervised domain adaptation framework tailored for nighttime VRU detection. TripleA includes triple alignment. First, it aligns daytime and nighttime images to generate synthetic nighttime images, which are then enhanced for illumination and noise. To remove noise, we introduce an illumination difference-aware denoising network, incorporating a novel pseudo-supervised attention to achieve pixel-wise noise distribution alignment. This alignment is driven by pseudo-ground truth generated through a carefully designed exchange-recombination strategy, facilitating self-supervised training of the denoising network. Additionally, we introduce degradation alignment to ensure domain-invariant degradation encoding, which enhances the network's robustness for real-world nighttime images. Extensive experiments demonstrate the effectiveness of our framework for nighttime VRU detection, all without the need for annotated nighttime data.**

***Index Terms*—Vulnerable road user detection, unsupervised domain adaptation, image enhancement, low-light, denoising.**

I. Introduction

VULNERABLE road users (VRUs) refer to non-motorized individuals on roadways, such as cyclists, pedestrians, and motorcyclists, who are particularly susceptible to transportation accidents [1], [2]. The detection of VRUs is crucial in various applications, including smart cities [3], intelligent transportation systems [4], and safety surveillance [5]. The recent boom in deep learning has led to great success in data-driven VRU detection methods [6], [7]. Although achievements have been made in related fields, the challenge persists in VRU detection under adverse weather, especially in nighttime VRU detection.

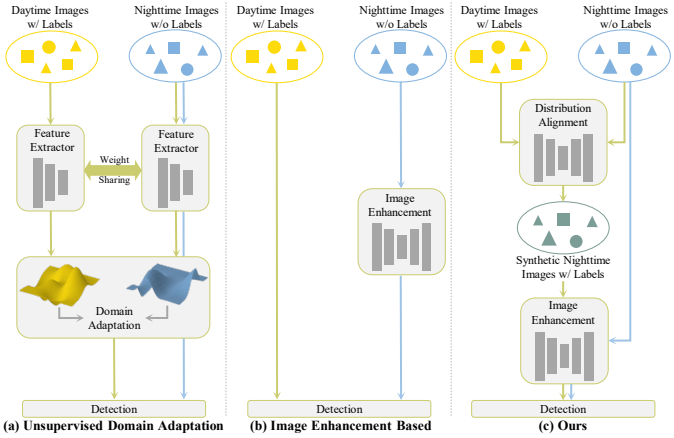


Fig. 1. Illustration of the paradigms for nighttime VRU detection without annotations. The green and blue arrows represent the training and testing data flows, respectively. (a) Unsupervised domain adaptation methods apply different alignment strategies to adapt the cross-domain distribution. (b) The nighttime images are enhanced, allowing the detectors trained on the daytime domain to be directly applied. (c) Our proposed framework aligns daytime and nighttime data distributions and removes degradation from nighttime images, allowing labeled daytime data to be fully utilized and machine vision to be free from the impairments of nighttime degradation.

Numerous state-of-the-art methods [8]–[10] have been proposed to improve detection performance under low-light conditions. Most of them are data-driven and depend on annotations of nighttime images. However, due to the poor quality of nighttime images, annotating them is labor-intensive and time-consuming, which significantly impedes the practical application of detection methods for nighttime VRUs.

Unsupervised domain adaptation (UDA) offers a promising approach to reduce dependence on nighttime annotations, as illustrated in Fig. 1(a). UDA techniques enable the adaptation of detection models trained on annotated daytime source domain data to the unlabeled nighttime target domain by learning domain-invariant features. For example, a trainable color-invariant convolution [11] transforms input images into edge representations that are independent of illumination and color. Other approaches [12], [13] focus on image- or instance-level feature alignment to minimize domain discrepancies. However, most of these methods focus mainly on semantic gaps between daytime and nighttime domains, often overlooking pixel-level gaps caused by nighttime degradation, such as pixel intensity and random noise. These unaddressed discrepancies can compromise the effectiveness of feature extraction and alignment

This work is supported by the Shanxi Provincial Science and Technology Major Special Project (202201150401020), Key R&D Project of China Metallurgical Group Corporation (2342G25S29), National Natural Science Foundation of China (42090012), Guangxi Science and Technology Plan Project (Guike 2021AB30019), Zhuhai Industry-University-Research Cooperation Project (ZH22017001210098PWC), Guangxi Key Laboratory of Spatial Information and Surveying and Mapping Fund Project (21-238-21-01), National Natural Science Foundation of China (62401410), The Fundamental Research Program of Shanxi Province (202203021222427). (Corresponding author: Zhenfeng Shao)

Yuankun Wang, Zhenfeng Shao, Yu Wang, Yulin Ding, and Gui Cheng are with the State Key Laboratory of information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: ykwang@whu.edu.cn; shaozhenfeng@whu.edu.cn; wy2022@whu.edu.cn; dingyulin@whu.edu.cn; chenggui@whu.edu.cn).

Jiaming Wang is with the Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430200, China (e-mail: wjmecho@whu.edu.cn).

strategies.

To address nighttime degradation, a straightforward approach involves enhancing nighttime images to reduce the pixel-level gaps relative to daytime data, thereby enabling detectors trained on daytime images to generalize effectively to nighttime scenes, as illustrated in Fig. 1(b). State-of-the-art enhancement methods can be categorized into three types: supervised, unsupervised, and self-supervised approaches. Supervised methods [14], [15] typically require paired low-light and normal-light images. However, obtaining ground truth for nighttime images is challenging due to the inherent degradations associated with nighttime scenarios. As a result, unsupervised and self-supervised methods, which do not rely on ground truth, have been proposed. These methods enhance images through generative method [16], high-order curve adjustment [17], and concave curve alignment [18]. However, these methods often struggle to suppress noise introduced by increased brightness during enhancement. Although these methods yield notable improvements in visual quality and image evaluation metrics, their enhancements do not reliably translate into performance gains for downstream detection tasks and may, in certain instances, even hinder detection accuracy.

As the methods mentioned above cannot address both semantic and pixel-level gaps between daytime and nighttime images, we propose TripleA, an unsupervised domain adaptation framework that incorporates triple alignment: distribution alignment, noise distribution alignment, and degradation alignment. The distribution alignment bridges the semantic gap by transferring the style of daytime images to nighttime scenes using a generative model with a high-frequency consistency constraint (see the upper part of Fig. 1(c)). The synthetic nighttime images generated in this way suffer from low illumination and noise, which impair machine vision and require image enhancement (see the lower part of Fig. 1(c)). Concretely, a physical model decouples illumination and noise for independent enhancement, with illumination first brightened and the noise subsequently suppressed by our illumination difference-aware denoising network. Within this network, the second alignment resolves pixel-level discrepancies in noise distribution between daytime and synthetic nighttime images via pseudo-supervised attention (PSA). This is achieved by using an exchange-recombination strategy to generate pseudo-ground truth, which provides pixel-wise references for self-supervised denoising. Degradation alignment enforces the denoising network to learn domain-invariant degradation features, thus improving its robustness on real-world nighttime images. The unified TripleA framework integrates with downstream detectors, enabling nighttime VRU detection without the need for annotated nighttime data. Extensive experiments on the KAIST Multispectral Pedestrian Detection Benchmark [19] and the EuroCity Persons [20] dataset demonstrate the superiority of TripleA. Further evaluation on two additional real-world test sets highlights its generalizability across diverse nighttime environments.

- We propose TripleA, an unsupervised domain adaptation framework for nighttime VRU detection that eliminates the need for nighttime annotations. Comprehensive experiments demonstrate its superiority over state-of-the-art methods.
- We design an illumination difference-aware denoising network that prioritizes denoising in regions with large illumination boost, thereby efficiently eliminating noise in nighttime images.
- We introduce pseudo-supervised attention, in which pseudo-ground truth generated via a novel exchange-recombination strategy provides a pixel-wise reference for the denoising network, enabling self-supervised training.

II. Related Work

A. Learning From Synthetic Data

Given the reliance of deep learning models on data, they typically excel when the training (source) and testing (target) domains share similar distributions. However, their performance can decline markedly under dissimilar distributions. To enhance model robustness, data augmentation techniques such as geometric transformations and color space adjustments are used to diversify training data. Despite these efforts, data augmentation may not bridge significant domain gaps effectively. To tackle this, some strategies include creating an intermediate domain to bridge the gap between the source and target domains [21], [22], while others adjust the data between the domains to align distributions [23], [24]. This study employs distribution alignment between the labeled daytime and unlabeled nighttime domains, generating a synthetic labeled nighttime domain to circumvent the need for annotating low-quality nighttime images.

B. Image Enhancement

The poor quality of nighttime images mainly lies in two aspects: low illumination and severe noise. Numerous methods have been proposed to brighten low-light images. Naive methods like histogram equalization and its variants [25] improve image visibility by redistributing the value distribution of the histogram. However, they often result in over-enhancement due to the neglect of the spatial distribution of illumination. There are also variational methods built upon imaging models, such as the atmospheric scattering model [26], [27] and the Retinex model. The Retinex model assumes an image I is made up of an illumination component L and a clean reflectance component R:

$$I = L \times R. \tag{1}$$

Many Retinex-based methods [28]–[30] first estimate illumination map L and then directly take I/L as light-enhanced version. These variational methods typically require lengthy inference times as each input necessitates a unique model solution. In recent years, the rapid advancement of deep learning has led to the widespread adoption of neural networks for image enhancement. These deep learning-based approaches can be categorized into supervised [14], [15], [31], [32], unsupervised [33]–[35], and zero-shot learning methods [17], [18], [36]. Representative supervised methods such as Retinex-Net [14], KinD [37], and Retinexformer [15]

are grounded in Retinex theory and are trained using large-scale paired datasets. Although these methods achieve superior performance, they depend heavily on paired training samples (a dark image paired with its well-lit counterpart), which are challenging and costly to collect in real-world scenarios. To circumvent the need for paired data, EnlightenGAN [16] utilizes unpaired training samples, where the scenes in low-light and well-lit images do not correspond. Additionally, to eliminate the requirement for well-lit images, the lightweight zero-shot learning method Zero-DCE [17], [36] treats image enhancement as a curve mapping problem.

Numerous aforementioned methods take image enhancement solely as a brightening issue, often overlooking the presence of noise inherent in low-light conditions. To address noise N, some strategies modify Eq. (1) as follows:

$$I = L \times R + N, \tag{2}$$

$$I = L \times (R + \frac{N}{L}) = L \times R', \tag{3}$$

where R' represents the noisy reflectance map. Consequently, these approaches simultaneously address illumination enhancement and noise suppression [15], [29], [30]. Specifically, JED [29] applies spatial smoothness to the reflectance map and addresses enhancement and denoising through a sequential decomposition model. Retinexformer [15], which includes an illumination estimator and a corruption restorer, is designed to brighten images and remove corruptions in a single stage. While these methods represent the state-of-the-art according to image quality assessment metrics, they do not necessarily benefit, and may even impair, downstream high-level vision tasks.

C. Vulnerable Road User Detection

A number of benchmarks for VRU detection have been established [38], [39]. However, most of them overlook night scenes where VRUs face increased danger. To address this gap, benchmarks [19], [40], [41] include night data, highlighting the challenges of nighttime scenes. While the majority of state-of-the-art methods [42], [43] utilize multi-modal data, such as thermal and gated images, and employ data fusion techniques to significantly enhance detection performance, the cost of these sensors can be prohibitive. Alternatively, some methods rely solely on RGB images. For instance, a detector named FATNet [44] improves nighttime features. Additionally, background information integrates into the channel attention mechanism [45] to expand the difference between low-illumination nighttime pedestrian features and the background. However, these methods depend on nighttime annotations, which are labor-intensive and time-consuming to obtain due to the inherently poor quality of nighttime images.

D. Dark Object Detection

Despite remarkable progress in object detection on high-quality images, performance often deteriorates, sometimes drastically, in challenging dark scenes. To address this, image enhancement prior to detection is an intuitive approach. ED-TwinsNet [46] jointly optimizes a learnable enhancement module with the detector. IA-YOLO [47] incorporates a differential image processing module before the detector, applying a sequence of image filters to improve quality. DENet [48], based on a Laplacian pyramid and cascaded with the detector, adaptively enhances and denoises images in a detection-driven manner. While these methods partially address dark object detection, they all require annotations of dark images. To eliminate this dependency, MAET [49] introduces a low-illumination-degrading pipeline that corrupts normal-light images, simultaneously estimating degrading parameters and detecting objects within a multitask learning framework. Another effective strategy is UDA, which uses well-lit images with labels as the source domain and unlabelled dark images as the target domain. Progressive DA [22] establishes an intermediate domain between the source and target, enabling progressive adaptation of features from the source to the target domain. HLA-Face [50] employs a bidirectional adaptation framework for dark face detection, adapting both low-level and high-level features. While some methods [12], [13] are not initially designed for dark object detection, their underlying principles are applicable. However, our experiments show that UDA methods underperform when the quality of dark images is poor.

III. Method

A. Overview

Let $\mathcal{D}$ represent the daytime source domain, where $\mathcal{D} = \{I^d, y\}$, with I^d denoting the daytime image and y the detection label. The nighttime target domain is indicated by $\mathcal{N}$, where $\mathcal{N} = \{I^n\}$, as target labels are unavailable.

The overall framework of TripleA is illustrated in Fig. 2 and the training procedure is shown in Algorithm 1, comprising distribution alignment and image enhancement. For distribution alignment, a CycleGAN [51] is adopted to learn from both $\mathcal{D}$ and $\mathcal{N}$. It then aligns the distribution of I^d with $\mathcal{N}$, resulting in a synthetic nighttime image $I^s \in \mathcal{S}$.

In the image enhancement component, I^s is decomposed into an illumination map L and a noisy reflectance map R' based on Retinex theory, which assumes that L is locally smooth and continuous. Here, L is derived by averaging the image channels and applying a Gaussian filter $G_\sigma(\cdot)$. According to Eq. (3), R' is calculated as I^s/L.

The illumination map L is enhanced using an illumination brightening network (section III-B1), while a denoising network suppresses noise in R' (section III-B2). To account for the interaction between illumination and noise, the denoising process is guided by the illumination difference map ΔL. Additionally, PSA (section III-B4) is incorporated to align the noise distribution and provide pseudo-supervision for denoising (section III-B3). Domain-invariant degradation alignment is also applied to extract robust degradation features, improving the denoising network's robustness (section III-B5).

Finally, an enhanced nighttime image I^s_{ehc} is obtained. A downstream detector is then trained using the enhanced image and its detection label (I^s_{ehc}, y).

B. Image Enhancement

1) Illumination Brightening: To improve the visibility of a nighttime image, we enhance its brightness by brightening

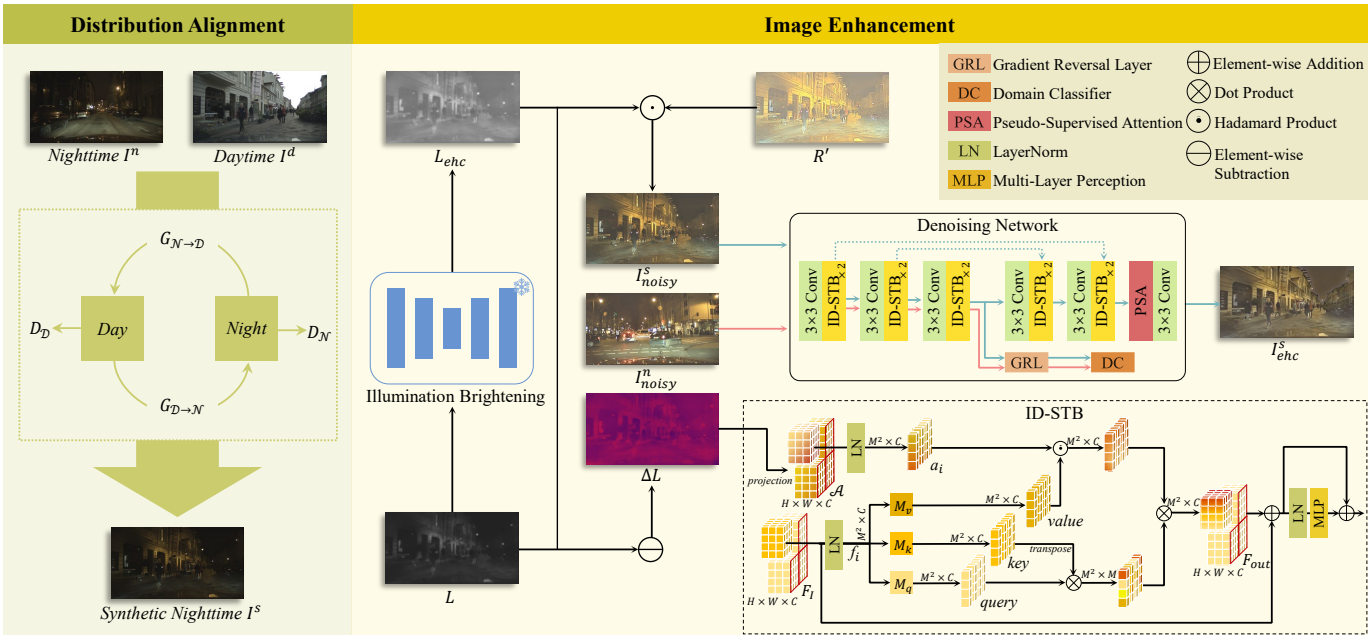


Fig. 2. The overall framework of TripleA. In distribution alignment, the daytime and nighttime data distributions are aligned to obtain a synthetic nighttime domain. In the enhancement phase, L and R' are the decomposed illumination and reflectance maps of the synthetic nighttime image I^s. L is enhanced by an illumination brightening network, and the noise in I^s_{noisy} is suppressed by the denoising network. An Illumination Difference-Aware Shifted Window Transformer Block (ID-STB) is designed to focus on denoising the areas with a large illumination boost. The PSA is integrated into the network to enable the self-supervised learning. A brightened real-world nighttime image I^n_{noisy} is involved in performing degradation alignment implemented by a gradient reversal layer and a domain classifier.

the decomposed illumination map L. Given that L is smooth and continuous, it does not contain any details of the original image but solely captures the ambient light within the scene. It is straightforward to adjust such a homogeneous map, and our primary emphasis lies not in designing a cutting-edge network architecture for illumination enhancement. Thus, we adopt an off-the-shelf high-order curve mapping approach [17] to brighten the illumination map L because of its efficiency and simplicity. The brightening can be expressed as:

$$L_n(p) = L_{n-1}(p) + A_{n-1}(p)L_{n-1}(p)(1 - L_{n-1}(p)), \quad (4)$$

where p refers position coordinates of illumination map, n is the number of iterations, A_{n-1} denotes element-wise parameter for the n-th order of the estimated curve.

2) Illumination Difference-Aware Denoising Network: The architecture of the denoising network is depicted in Fig. 2. It receives a brightened noisy image as its input, which can be expressed as:

$$I_{noisy} = L_{ehc}(R + \frac{N}{L}) = L_{ehc}R + \frac{L_{ehc}}{L}N, \quad (5)$$

$$I_{noisy} = I_{ehc} + \frac{L_{ehc}}{L}N. \quad (6)$$

We aim to restore I_{ehc} and remove the degradation term $\frac{L_{ehc}}{L}N$. From Eq. (6), it can be seen that the noise is coupled with the scale of the brightened illumination map L_{ehc} to the initial illumination map L. It is also in accordance with the real-world observation: when a nighttime image is brightened, regions with a large illumination boost reveal more noise than regions with a small illumination boost. Thus, it is vital for the denoising network to be aware of the spatial distribution of illumination improvement.

An intuitive idea is taking the term $\frac{L_{ehc}}{L}$ as a guide map to gear the denoising learning focus on particular regions, whereas the division operation can result in overflow values which is detrimental to stable learning. Hence, we employ the illumination difference map $\Delta L = |L_{ehc} - L|$ as the guide.

We introduce the Illumination Difference-Aware Shifted Window Transformer Block (ID-STB), which serves as the core component of the denoising network. The ID-STB leverages the Swin-Transformer block [52] as its base structure, suitable for addressing the local correlation and non-stationarity of noise that necessitates both local and long-range modelings. To guide the modeling, the illumination difference map ΔL is projected into an attention feature space $\mathcal{A} \in \mathbb{R}^{H\times W\times C}$, which subsequently guides the computation of multi-head self-attention (MSA) within the ID-STB.

The schematic diagram of the ID-STB is depicted in the lower right corner of Fig. 2. The inputs to the ID-STB include the image feature map $F_I \in \mathbb{R}^{H\times W\times C}$ and the projected feature of the illumination difference map $\mathcal{A}$. These inputs are first normalized using a LayerNorm and then participate in the self-attention computation. Subsequently, they pass through another LayerNorm layer followed by a Multi-Layer Perceptron (MLP). The input data are partitioned into $\frac{HW}{M^2}$ windows of $M\times M$ size and reshaped into $\frac{HW}{M^2}\times M^2\times C$. The self-attention is computed independently within each window using the local window features $f \in \mathbb{R}^{M^2\times C}$ and $a \in \mathbb{R}^{M^2\times C}$. The attention matrices $query$ (q), key (k), and $value$ (v) for single-head attention within a window are computed as

Algorithm 1: Training Procedure of TripleA

Procedure *Distribution Alignment*
  for M_1 *epochs* **do**
    for m_1 *steps* **do**
      Draw a batch of images $\{I^d_{(1)}, ..., I^d_{(b_1)}\}$, $\{I^n_{(1)}, ..., I^n_{(b_1)}\}$ from domain $\mathcal{D}$ and $\mathcal{N}$, respectively;
      Input $\{I^d_{(i)}\}^{b1}_{i=1}$ and $\{I^n_{(i)}\}^{b1}_{i=1}$ into $G_{\mathcal{D}\to\mathcal{N}}$, $G_{\mathcal{N}\to\mathcal{D}}$, $D_{\mathcal{D}}$, and $D_{\mathcal{N}}$ to learn the distribution of $\mathcal{D}$ and $\mathcal{N}$;
      Update the parameters of $G_{\mathcal{D}\to\mathcal{N}}$, $G_{\mathcal{N}\to\mathcal{D}}$, $D_{\mathcal{D}}$, and $D_{\mathcal{N}}$ by minimizing $\mathcal{L}_{da}$ defined in Eq. (12);
    end
  end
  Obtain I^s through unidirectional mapping $I^s = G_{\mathcal{D}\to\mathcal{N}}(I^d)$;
end
Procedure *Image Enhancement*
  for M_2 *epochs* **do**
    for m_2 *steps* **do**
      Draw a batch of images $\{I^s_{(1)}, ..., I^s_{(b_2)}\}$, $\{I^n_{(1)}, ..., I^n_{(b_2)}\}$ from domain S and N, respectively;
      for $I \in \{I^s, I^n\}$ **do**
        Decompose I into R' and L;
        Feed L into illumination brightening network to generate L_{ehc} and obtain $\Delta L = |L_{ehc} - L|$;
        Obtain I_{noisy} with Eq. (5);
        if I *is* I^s **then**
          $I^s_{noisy} \leftarrow I_{noisy}, \Delta L^s \leftarrow \Delta L$;
        else
          $I^n_{noisy} \leftarrow I_{noisy}, \Delta L^n \leftarrow \Delta L$;
        end
      end
      `// Use the Denoising Network's Encoder and Decoder`
      Obtain $Fea^s_{mid} = Encoder(I^s_{noisy}, \Delta L^s)$ and $I^s_{ehc} = Decoder(Fea^s_{mid})$;
      Obtain $Fea^n_{mid} = Encoder(I^n_{noisy}, \Delta L^n)$;
      Input Fea^s_{mid} and Fea^n_{mid} to the domain classifier;
      Update the parameters of the denoising network by minimizing $\mathcal{L}_{dns}$ defined in Eq. (15);
    end
  end
end

follows:

$$q = f_i M_q, \quad k = f_i M_k, \quad v = f_i M_v, \tag{7}$$

where $f_i \in \mathbb{R}^{M^2 \times \frac{C}{h}}$ is the i-th head after f is split into h heads. M_q, M_k, M_v are projection matrices. To achieve the illumination difference guide, attention feature a is also split into h heads to obtain $a_i \in \mathbb{R}^{M^2 \times \frac{C}{h}}$, which spatially weights the $value$ (v):

$$f_{out} = (v \odot a_i) \otimes softmax(q \otimes k^T / \sqrt{s} + b), \tag{8}$$

where $\odot$ and $\otimes$ are Hadamard product and dot product, respectively. k^T is the transpose of k. s is a scaling parameter. b is the learnable relative positional encoding.

3) Exchange-Recombination Strategy: Unlike low-light images, the intrinsic degradation associated with nighttime scenes imposes significant challenges in acquiring a bright, clean ground truth image that could supervise the denoising network. In response to this limitation, and benefiting from the distribution alignment stage in our framework, we propose an exchange-recombination strategy. It creates a reference for the denoising network which enables self-supervised learning.

The entire process is illustrated in Fig. 3(a). Specifically, through distribution alignment, a daytime image I^d and its synthetic nighttime counterpart I^s are obtained. Despite the significant differences in texture, color, contrast, spatial frequency, and other characteristics due to their distinct modalities, the objects contained within I^d and I^s remain the same. Moreover, I^d is typically less affected by noise compared to I^s. Hence, we align the noise distribution of I^s with that of I^d to mitigate night-specific noise. In this regard, the reflectance components of I^s and I^d are exchanged, and the daytime reflectance R^d is recombined with the enhanced illumination L^s_{ehc} to create a pseudo-ground truth:

$$I_{sGT} = L^s_{ehc} \odot avg(R^d), \tag{9}$$

where $\odot$ signifies the Hadamard product, $avg(\cdot)$ denotes the channel-wise average. $I_{sGT} \in \mathbb{R}^{H \times W \times 1}$ is less influenced by noise and irrelevant to color, making it an ideal reference for the noise distribution. This facilitates the denoising network's ability to use I_{sGT} as an explicit supervision, guiding the network's output towards noise reduction while maintaining the integrity of the image structure.

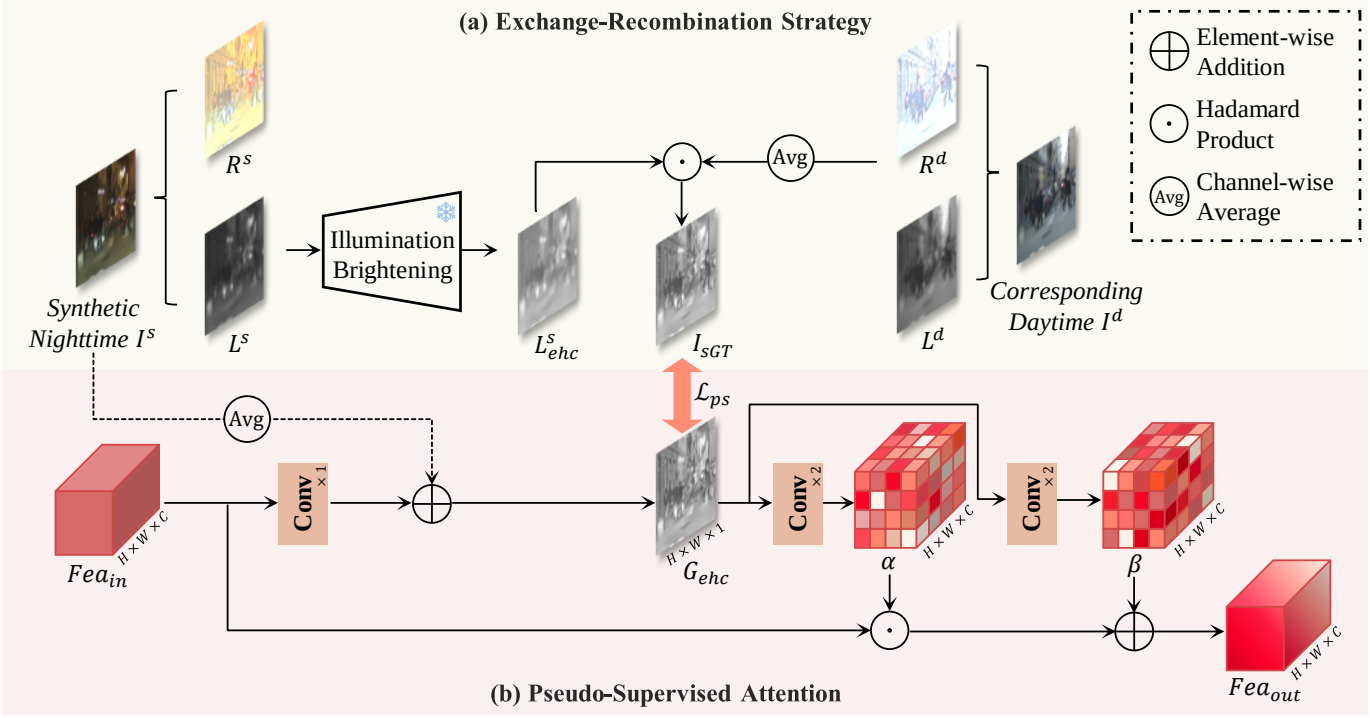


Fig. 3. Illustration of the (a) exchange-recombination strategy and (b) PSA. The pseudo ground truth I_{sGT} formulated by exchange-recombination strategy provides a pixel-wise reference for the intermediate output G_{ehc}. $\mathcal{L}_{ps}$ denotes the pseudo supervision loss.

4) Psuedo-Supervised Attention: The PSA is presented to perform pseudo-supervision by aligning the noise distribution between the denoised output and the pseudo-ground truth I_{sGT}. It is detailed in the schematic diagram shown in Fig. 3(b). The intermediate output G_{ehc} is extracted to be supervised, as this not only refines the final output features, but also avoids color deviation. Moreover, inspired by [53], the PSA utilizes an affine transformation to recalibrate the output features. Initially, PSA learns a pair of modulation parameters (α, β), which are conditioned on G_{ehc}. These parameters are then employed to adjust the output features, which can be written as:

$$\alpha = \mathcal{M}_1(Fea_{in}|G_{ehc}), \quad \beta = \mathcal{M}_2(Fea_{in}|G_{ehc}), \tag{10}$$

$$Fea_{out} = \alpha Fea_{in} + \beta, \tag{11}$$

where Fea_{in} and Fea_{out} are input and output features, respectively. $\mathcal{M}_1$ and $\mathcal{M}_2$ are learnable mapping functions conditioned on G_{ehc}.

5) Degradation Alignment: The denoising network trained on synthetic nighttime images is expected to perform well on real-world nighttime images. To further close the gap between the two domains, we propose to align the degradation of them, so that the encoder can extract domain-invariant degradation. To this end, a domain classifier is introduced between the encoder and the decoder to classify the domain of the encoded degradation features. The encoder can focus on extracting domain-robust degradation features to fool the domain classifier. The adversarial learning between the domain classifier and the encoder is achieved by a gradient reversal layer (GRL) [54]. Specifically, the loss function of the domain classifier is minimized in the forward pass, while it is maximized by negating the gradients flowing through the GRL in the backward pass. In this way, the degradation of different domains is aligned thereby improving the robustness of the denoising network.

C. Loss Functions

1) Distribution Alignment Loss Functions: The full objective of the distribution alignment can be written as:

$$\mathcal{L}_{da} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{hfc}\mathcal{L}_{hfc}, \tag{12}$$

where $\mathcal{L}_{hfc}$ denotes our proposed high-frequency consistency loss, $\mathcal{L}_{adv}$ and $\mathcal{L}_{cyc}$ represent the adversarial loss and cycle-consistency loss, respectively, adopting their exact forms from [51]. The hyperparameters λ_{adv}, λ_{cyc}, and λ_{hfc} are employed to balance the terms.

High-Frequency Consistency Loss. It has been observed that the generative network exhibits a "spectral bias" towards low-frequency information [55], resulting in blurry edges in synthetic images. Therefore, we introduce a high-frequency consistency loss to encourage the network to focus on high-frequency details, such as fine edges and rich textures, which are crucial for VRU detection. Specifically, we integrate edge extraction into the cycle consistency constraint of the generative model. This ensures that ineffective information flow is minimized in bidirectional mapping (both day-to-night $G_{\mathcal{D}\to\mathcal{N}}$ and night-to-day $G_{\mathcal{N}\to\mathcal{D}}$) and helps the model become more aware of high-frequency details of objects. We use the L1 norm to penalize the edge differences between real images and their synthetic counterparts, as it is robust to noise. The high-frequency consistency loss is formulated as follows:

$$\mathcal{L}_{hfc}(G_{\mathcal{D}\to\mathcal{N}}, G_{\mathcal{N}\to\mathcal{D}}) = \mathbb{E}_{I^d\sim p_{\text{data}}(I^d)}\|S(G_{\mathcal{N}\to\mathcal{D}}(G_{\mathcal{D}\to\mathcal{N}}(I^d))) - S(I^d)\|_1 + \mathbb{E}_{I^n\sim p_{\text{data}}(I^n)}\|S(G_{\mathcal{D}\to\mathcal{N}}(G_{\mathcal{N}\to\mathcal{D}}(I^n))) - S(I^n)\|_1, \tag{13}$$

where $S(\cdot)$ denotes the edge extraction operation. We use the Sobel operator for $S(\cdot)$ due to its robustness to noise and its differentiability. The edge extraction from an image I can be expressed as:

$$S(I) = \sqrt{(s_x * I)^2 + (s_y * I)^2}, \tag{14}$$

where s_x is the horizontal kernel, s_y is the vertical kernel, and $*$ denotes the convolution operation.

2) Image Denoising Loss Functions: To optimize the denoising network, we design the following loss functions, including the pseudo supervision loss, color consistency loss, reconstruction loss, perceptual loss, and domain classification loss.

$$\mathcal{L}_{dns} = \lambda_{ps}\mathcal{L}_{ps} + \lambda_{col}\mathcal{L}_{col} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{dc}\mathcal{L}_{dc}, \tag{15}$$

where $\lambda_{ps}, \lambda_{col}, \lambda_{rec}, \lambda_{per}, \lambda_{dc}$ are hyperparameters.

Pseudo Supervision Loss. It utilizes the formulated pseudo-ground truth as a reference to supervise the intermediate output of the denoising network:

$$\mathcal{L}_{ps} = ||I_{sGT} - G_{ehc}||_2^2. \tag{16}$$

Color Consistency Loss. It guarantees that the color of the image is consistent before and after denoising.

$$\mathcal{L}_{col} = \sum_c (U_c - V_c)^2, c = \{R, G, B\}, \tag{17}$$

where U_c and V_c denote the average intensity value of c channel in I^s_{ehc} and I^s_{noisy}.

Reconstruction Loss. It ensures that the denoised image is not dramatically different from the input.

$$\mathcal{L}_{rec} = ||I^s_{noisy} - I^s_{ehc}||_2^2. \tag{18}$$

Perceptual Loss. Considering the consistency of the semantic feature between the input I^s_{noisy} and output I^s_{ehc} of the denoising network, we constrain it by:

$$\mathcal{L}_{per} = ||VGG_i(I^s_{noisy}) - VGG_i(I^s_{ehc})||_2^2, \tag{19}$$

where VGG_i is the operation of extracting features from the i-th layer of a pre-trained VGG-19 [56].

Domain Classification Loss. It utilizes a binary cross entropy loss to penalize the predicted domain class p based on the real domain label d of input features, where features from real-world night domain $\mathcal{N}$ are given the label $d = 1$ and the features from synthetic night domain $\mathcal{S}$ receive label $d = 0$.

$$\mathcal{L}_{dc} = -d\log p - (1-d)\log(1-p). \tag{20}$$

IV. Experiments

A. Experimental Details

Datasets. In this work, we conduct a comprehensive evaluation and comparison of state-of-the-art methods using two large-scale benchmarks for VRU detection. These benchmarks consist of images captured during both daytime and nighttime. For our study, we exclusively use labeled daytime training sets and unlabeled nighttime training sets. Annotations from the nighttime test set are used solely for evaluation purposes during the testing phase.

KAIST Multispectral Pedestrian Detection Benchmark [19], acquired within typical traffic scenes, features well-aligned color-thermal image pairs. In this paper, we selectively extracted the visible images from the original dataset, resulting in a collection that encompasses both daytime and nighttime visible imagery, each with a spatial resolution of 640×512 pixels. Owing to numerous problematic annotations in the original dataset, we adopted the sanitized training annotations provided by [57] and the improved test annotations offered by [58]. Our training set thus consists of 7601 images, including 4755 captured during the daytime and 2846 during nighttime, alongside an additional 797 nighttime images in our test set. In the sanitized version [57] of the dataset, both the 'person' and 'cyclist' categories have been uniformly labeled as 'person'. This alteration stems from the challenges faced by human annotators when distinguishing between pedestrians and cyclists under adverse lighting conditions or when working with low-resolution imagery. Consequently, we only use the 'person' class in this study.

EuroCity Persons (ECP) [20] dataset, collected across 32 cities within 12 European countries, offers a substantial repository of detailed annotations featuring a diverse range of pedestrians, cyclists, and other riders in urban traffic scenarios. This dataset contains over 238,200 annotated instances across more than 47,300 images, each with a spatial resolution of 1920×1024 pixels. In this study, we designate the 'pedestrian' and 'rider' classes as our detection targets. We follow the official training set splitting and use the original validation set as our test dataset, due to the unavailability of publicly accessible annotations for the test set. As a result, our dataset includes 28,114 training images, with 23,892 captured during the daytime and 4,222 during nighttime, plus an additional 770 nighttime images in our test set.

Training Settings. In the distribution alignment phase, the CycleGAN processes daytime I^d and nighttime I^n images as inputs during training. After training, it performs a unidirectional translation to produce a synthetic nighttime image, I^s.

Our implementation is based on PyTorch. For the CycleGAN, we apply a batch size of 4 and use the Adam optimizer [59] with an initial learning rate of 2×10^{-4}. We set the training epochs to 100, with the learning rate halving after the first 50 epochs. In Eq. (12), the loss weights λ_{adv}, λ_{cyc} are set to 10 and 0.5, respectively, following the default implementation [51]. The weight λ_{hfc} is empirically set to 5 to maintain balance among the loss terms.

For the denoising network, a total of 900 synthetic nighttime images paired with their corresponding daytime counterparts

Fig. 4. Comparison with enhancement-based methods on the KAIST dataset. (a) Input nighttime image and ground truth bounding boxes. (b)-(n) Detection results on images enhanced using enhancement-based methods. (o) Our result. * indicates the retrained version of the deep learning method.

TABLE I
THE SETTINGS OF THE ECP TEST SET.

Setting	Height (px)	Occlusion (%)
reasonable	[40, inf]	[0, 40]
small	[30, 60]	[0, 40]
occluded	[40, inf]	[40, 80]

are used for training. During the training process, each image is cropped into 128×128 patches. The network is optimized using the Adam optimizer [59] for 100 epochs. The initial learning rate of 2×10^{-4} is gradually reduced to 1×10^{-7} following a cosine annealing schedule [60]. A batch size of 8 is employed. The hyperparameters in Eq. (15) are empirically set as follows: $\lambda_{ps} = 1 \times 10^2$, $\lambda_{col} = 10$, $\lambda_{rec} = 5 \times 10^{-5}$, $\lambda_{per} = 1 \times 10^{-1}$, $\lambda_{dc} = 1$. These values are chosen to ensure that each loss term contributes comparably, maintaining a balance among the constraints.

Evaluation Settings. We employ the Log Average Miss Rate (LAMR), a metric commonly used in pedestrian detection, and mean Average Precision (mAP) for quantitative assessments. In all our evaluations, we adhere to the common practice of using an Intersection over Union (IoU) threshold of 0.5, appropriate due to the high non-rigidness of VRUs. On the ECP dataset, we report numbers for different occlusion levels: "reasonable", "small", and "occluded", as defined in Table I. Additionally, we also provide the numbers for the "all" test case.

B. Comparison Results

For a comprehensive evaluation, we compare our framework with state-of-the-art methods across multiple categories. We employ four commonly-used detectors (Cascade R-CNN [61], Faster R-CNN [62], YOLOv5s [63], and YOLOv8s [64]) trained on daytime images and directly tested on nighttime images. The most generalizable detector is then selected as the base detector for each dataset. In the enhancement-based category, the base detector is first trained on daytime images and then tested on nighttime images enhanced by various methods (ZeroDCE [17], EnlightenGAN [16], RetinexNet [14], JED [29], Retinexformer [15], LIME [28], SNR [65], KinD [37], KinD++ [66], GLARE [67], CIDNet [68]). We

TABLE II
COMPARISON RESULTS ON THE KAIST DATASET.

Category	Method	LAMR(%) ↓	mAP(%) ↑
Generalization	Cascade R-CNN [61]	77.1	43.7
	Faster R-CNN [62]	85.2	41.1
	YOLOv5s [63]	89.5	51.3
	YOLOv8s [64]	92.6	47.5
Oracle	Fine-tuned Cascade R-CNN [61]	61.7	52.6
Enhancement Based (with Cascade R-CNN)	ZeroDCE [17]	85.0	35.4
	ZeroDCE* [17]	82.7	42.2
	EnlightenGAN [16]	80.6	43.5
	EnlightenGAN* [16]	74.7	41.0
	RetinexNet [14]	87.4	34.5
	JED [29]	80.0	41.5
	Retinexformer [15]	82.7	41.5
	LIME [28]	82.6	36.3
	SNR [65]	79.8	38.5
	KinD [37]	83.8	38.5
	KinD++ [66]	84.3	36.7
	GLARE [67]	80.0	42.4
	CIDNet [68]	81.7	42.9
Unsupervised Domain Adaptation	EPM [12]	86.2	23.1
	SIGMA++ [13]	88.4	17.3
	ALDI++ [69]	69.5	56.2
	Simple SFOD [70]	88.8	15.1
	SNR† [65]	67.4	46.4
	GLARE† [67]	71.1	41.5
	Ours	64.2	49.2

* denotes retraining the enhancement methods with our real-world nighttime images. † represents the image enhancement with this method after going through the same distribution alignment step as in our framework. (Key: Best, Second Best, Third Best)

also compare UDA methods (EPM [12], SIGMA++ [13], ALDI++ [69], Simple SFOD [70]), with daytime as the source domain and nighttime as the target domain. Notably, to further demonstrate the effectiveness of the proposed enhancement method in our framework, we enhance synthetic nighttime images using comparative enhancement methods after performing the distribution alignment process of our framework, and then train the detector. We select the two most effective enhancement methods for comparison based on the results in the enhancement-based categories, denoted by †. Furthermore, we provide an oracle where the detector is trained on daytime images and then fine-tuned on nighttime images, offering an upper limit for the results.

1) Evaluation on the KAIST Dataset: As depicted in Table II, Cascade R-CNN outperforms the other three detectors, achieving the lowest LAMR at 77.1%, and is therefore selected as the baseline detector.

All enhancement methods, except for EnlightenGAN*, result in decreased detection performance compared to the baseline. This decline can primarily be attributed to two factors. Firstly, most enhancement methods do not effectively improve nighttime image quality. For instance, GLARE fails to adequately brighten the image, as shown in Fig. 4(m); while RetinexNet, LIME, KinD, and KinD++ enhance brightness but simultaneously amplify noise (Fig. 4(f), (i), (k), (l)). Although JED (Fig. 4(g)) improves both brightness and noise reduction, it also blurs critical image details. In contrast, our method effectively eliminates noise in areas with significant illumination improvements, as demonstrated in the right part of Fig. 4(o), facilitated by the illumination difference map guiding the denoising network. Additionally, the method preserves object boundaries distinctly, due to the pixel-wise refinement implemented in the PSA.

Secondly, although image quality is ameliorated (as shown in Fig. 4(j)), the substantial gap between daytime and nighttime domains hinders detection performance. However, when the domain gap is reduced through the distribution alignment of our framework, SNR† and GLARE† show notable improvements in detection performance, with LAMR reductions of 12.4% and 8.9%, respectively, compared to their enhancement-only counterparts. This highlights the effectiveness of distribution alignment, further supported by the results from EnlightenGAN*, which translates nighttime images to the daytime domain. Our method achieves the best result, reducing LAMR to 64.2%, attributed to effective distribution alignment and superior enhancement, including noise elimination in dark areas and preservation of sharp contour details (Fig. 4(o)).

UDA methods typically use feature alignment, self-distillation, or a combination of both to address the domain shift between source and target domains. EPM and SIGMA++ primarily rely on feature alignment strategies; however, both fail to achieve optimal performance, indicating that pixel-level degradation impedes effective feature extraction and alignment. Simple SFOD and ALDI++ employ self-distillation strategies, with ALDI++ achieving the highest mAP due to its multi-task soft distillation approach, which enhances pseudo-label accuracy. However, its higher LAMR suggests that noise interferes with the detector's ability to recognize VRUs. In contrast, our method achieves the lowest LAMR, which directly correlates with fewer missed detections—an essential factor for ensuring the safety of VRUs.

2) Evaluation on the ECP Dataset: In contrast to the Cascade R-CNN's performance on the KAIST dataset, Table III shows that YOLOv8s exhibits superior generalization on the ECP dataset. Therefore, YOLOv8s is used as the baseline detector for the ECP dataset.

Similar to the KAIST dataset results, all enhancement-based methods, except for EnlightenGAN*, degrade detection performance. Some methods, such as RetinexNet (Fig. 5(f)) and KinD++ (Fig. 5(l)), cause over-enhancement, leading to oversaturation and unnatural artifacts. Others, like ZeroDCE (Fig. 5(b)) and its retrained counterpart (Fig. 5(c)), brighten the images while introducing a haze-like effect, which reduces contrast. In contrast, our enhancement result (Fig. 5(o)) improves brightness, minimizes artifacts, and preserves sharp details, leading to fewer missed detections and misclassifications. Although the enhanced result from EnlightenGAN* exhibits suboptimal visual quality, Table III shows that it reduces LAMR across all testing subsets compared to the baseline, highlighting the importance of aligning source and target domain distributions in machine vision.

Among UDA methods, EPM and SIGMA++ perform poorly on the 'small' test set, indicating that feature alignment for small targets is significantly hindered by degradation in nighttime images, leading to numerous false negatives. ALDI++ achieves a low LAMR in detecting riders, but this comes

Fig. 5. Comparison with enhancement-based methods on the ECP dataset. (a) Input nighttime image and ground truth bounding boxes. (b)-(n) Detection results on images enhanced using enhancement-based methods. * indicates the retrained version of the deep learning method. The pink bounding box represents the pedestrian class and the green bounding box represents the rider class.

at the expense of a low mAP, suggesting a high number of false positives. In contrast, results from JED†, GLARE†, and our method demonstrate that detection performance improves significantly when both domain distribution disparities and pixel-level degradation are addressed. Notably, our method not only achieves the lowest LAMR on several test settings, but also achieves a high mAP, indicating a balanced trade-off between precision and recall. This suggests that our denoising network effectively suppresses noise that interferes with the recognition and localization of VRUs.

C. Ablation Studies

In this section, we conduct a comprehensive ablation study on the ECP dataset to investigate the effectiveness of the proposed components.

High-Frequency Consistency Loss $\mathcal{L}_{hfc}$. To evaluate its effectiveness, we train the detector using synthetic nighttime images generated without this loss and assess its performance on real-world nighttime images, without applying image enhancement. As shown in Table IV, the improvement in detection performance with $\mathcal{L}_{hfc}$ demonstrates its ability to effectively reduce the domain gap between synthetic and real-world nighttime images.

Loss Functions of Image Denoising Network. In our illumination difference-aware denoising network, we individually ablate each loss function to assess its impact on the final detection accuracy. As shown in Table V, removing any of the loss functions increases the LAMR by approximately 1%, underscoring the importance of each constraint in the denoising process. Notably, excluding the domain classification loss results in the greatest increase in LAMR (around +2%), suggesting that aligning the encoded degradation features facilitates the extraction of real-world noise, thereby enhancing the quality of recovered nighttime images. These findings highlight the effectiveness of our proposed degradation alignment approach.

Illumination Brightening and Denoising Network. We compare the results of applying only illumination brightening and only image denoising. We clarify that when illumination brightening is not applied, the original illumination map L is used as a substitute to compute the ID-STB in the denoising network, as ΔL is unavailable. The experimental results are presented in Table VI. Compared to the baseline, employing only the denoising network reduces the LAMR, whereas using only the illumination brightening leads to an increase in the LAMR. Importantly, the combination of the two yields the most favorable outcomes, achieving a LAMR of 45.9% and

TABLE III
COMPARISON RESULTS ON THE ECP DATASET.

Category	Method	LAMR(reasonable) ↓		LAMR(small) ↓		LAMR(occluded) ↓		LAMR(all) ↓		mAP ↑	
		pedestrian	rider	pedestrian	rider	pedestrian	rider	pedestrian	rider	pedestrian	rider
Generalization	Cascade R-CNN [61]	38.5	59.6	69.8	93.6	69.4	53.8	51.4	63.5	57.3	38.2
	Faster R-CNN [62]	51.3	59.3	94.5	96.1	72.9	53.8	60.8	63.4	54.6	39.9
	YOLOv5s [63]	38.5	60.2	75.7	88.6	69.6	53.8	51.1	63.5	72.1	60.3
	YOLOv8s [64]	37.2	57.2	74.4	85.7	68.1	53.9	49.2	61.7	73.6	65.0
Oracle	Fine-tuned YOLOv8s [64]	32.7	48.3	62.7	65.7	63.7	53.8	45.2	52.5	77.6	71.1
Enhancement Based (with YOLOv8s)	ZeroDCE [17]	56.4	62.6	87.1	88.6	82.4	53.8	66.1	65.3	63.5	55.4
	ZeroDCE* [17]	46.3	59.6	81.0	82.9	74.5	53.8	57.3	62.9	69.1	56.5
	EnlightenGAN [16]	47.9	59.3	84.5	88.9	77.5	53.8	59.1	63.4	67.1	52.5
	EnlightenGAN* [16]	36.3	57.6	70.9	88.9	67.0	53.8	47.9	62.0	73.7	60.7
	RetinexNet [14]	53.1	61.1	82.9	83.8	81.5	53.8	63.4	64.2	64.9	53.1
	JED [29]	46.5	56.9	85.4	92.0	76.8	53.8	57.9	61.5	68.6	51.3
	Retinexformer [15]	45.5	59.6	79.1	88.6	74.9	53.8	56.7	63.6	69.4	60.5
	LIME [28]	48.3	59.8	82.4	91.4	79.7	53.8	59.8	63.8	67.4	54.6
	SNR [65]	49.7	59.8	86.7	92.1	77.0	53.8	60.4	63.3	66.4	53.4
	KinD [37]	44.2	65.1	76.3	86.3	76.1	53.8	55.9	67.3	68.9	50.0
	KinD++ [66]	51.3	63.6	81.7	88.6	81.0	53.8	62.1	66.8	65.2	48.6
	GLARE [67]	38.4	56.5	78.9	88.6	68.1	53.8	50.4	61.1	73.0	61.1
	CIDNet [68]	47.7	58.1	83.7	91.1	76.9	53.8	58.7	62.4	68.4	48.7
Unsupervised Domain Adaptation	EPM [12]	62.5	66.8	91.5	95.5	81.4	63.2	70.3	70.1	50.7	30.4
	SIGMA++ [13]	62.6	63.1	93.1	96.2	80.0	56.3	70.2	66.6	50.1	33.6
	ALDI++ [69]	39.6	50.7	75.7	76.0	69.5	50.3	51.7	54.9	68.3	49.5
	Simple SFOD [70]	87.0	80.0	96.8	100.0	95.4	69.2	90.3	81.4	22.7	15.1
	JED† [29]	28.9	53.1	63.7	83.2	61.2	53.8	41.5	57.7	77.1	63.8
	GLARE† [67]	27.8	51.6	61.7	82.9	57.0	53.8	39.9	55.6	77.9	68.1
	Ours	25.6	48.2	58.7	80.3	55.0	53.8	37.9	54.0	79.0	66.6

* denotes retraining the enhancement methods with real-world nighttime images. † represents the image enhancement with this method after going through the same distribution alignment step as in our framework. (Key: Best, Second Best, Third Best)

a mAP of 72.8%. This highlights that although illumination brightening enhances nighttime images, its exclusive use can impair performance by amplifying noise. However, the combined use of illumination brightening and denoising effectively suppresses the noise, demonstrating the denoising network's ability to mitigate the amplified artifacts of brightening. This emphasizes the importance of using both techniques together to achieve optimal nighttime image enhancement.

Illumination Difference Guide. We remove the illumination difference guide in ID-STB to verify its effectiveness. The quantitative results, presented in the first row of Table VII, show that the guide notably improves the detection of small targets, evidenced by a 2.2% reduction in LAMR(small). A visual comparison between the results with and without the guide (Fig. 6(c) and Fig. 6(f), respectively) reveals that, without the guide, small pedestrians in distant and low-light regions are missed. This highlights that it enhances the network's capacity to preserve subtle yet semantically meaningful features, which are crucial for accurately localizing small targets in low-light environments.

Pseudo-Supervised Attention. We jointly ablate PSA and the exchange-recombination strategy to evaluate their effectiveness, as the latter is specifically designed to support PSA. A comparison of the results in Fig. 6(d) and Fig. 6(f) reveals that the absence of PSA leads to increased noise, causing missed detections of pedestrians with less distinctive contours in the distance. Similarly, comparing the results in the second and fourth rows of Table VII demonstrates that introducing PSA enhances the detection performance of pedestrians and riders of all sizes. This finding indicates that PSA effectively refines feature representations and preserves more discriminative details.

Degradation Alignment. To verify the effectiveness of

TABLE IV
ABLATION OF HIGH-FREQUENCY CONSISTENCY LOSS

w/o L_{hfc}	w/ L_{hfc}	LAMR(%) ↓	mAP(%) ↑
✓		50.9	71.1
	✓	49.1	71.9

TABLE V
ABLATION OF LOSS FUNCTIONS IN IMAGE DENOISING NETWORK

Method	LAMR(%) ↓	mAP(%) ↑
Ours w/o $\mathcal{L}_{ps}$	46.8	72.9
Ours w/o $\mathcal{L}_{col}$	46.9	72.8
Ours w/o $\mathcal{L}_{rec}$	46.6	71.8
Ours w/o $\mathcal{L}_{per}$	46.8	73.8
Ours w/o $\mathcal{L}_{dc}$	47.6	72.4
Ours	45.9	72.8

TABLE VI
ABLATION OF ILLUMINATION BRIGHTENING AND DENOISING NETWORK.

Baseline	Illumination Brightening	Denoising Network	LAMR(%) ↓	mAP(%) ↑
✓			49.1	71.9
✓	✓		50.7	71.9
✓		✓	46.5	72.1
✓	✓	✓	45.9	72.8

TABLE VII
ABLATION OF THE KEY COMPONENTS OF DENOISING NETWORK. ID REFERS TO ILLUMINATION DIFFERENCE GUIDE, PSA REFERS TO PSEUDO-SUPERVISED ATTENTION, AND DA REFERS TO THE DEGRADATION ALIGNMENT.

ID	PSA	DA	LAMR(reasonable) ↓	LAMR(small) ↓	LAMR(occluded) ↓	LAMR(all) ↓	mAP ↑
	✓	✓	37.7	71.7	52.5	46.7	71.8
✓		✓	38.8	69.7	54.5	47.7	72.3
✓	✓		38.8	73.2	52.2	47.6	72.4
✓	✓	✓	36.9	69.5	54.4	45.9	72.8

TABLE VIII
IMPACT OF SYNTHETIC NIGHTTIME IMAGE QUALITY ON DENOISING NETWORK AND DETECTION PERFORMANCE

Experimental Configuration	**FID ↓**	**NIQE ↓**	**mAP (%) ↑**			**LAMR(all) (%) ↓**		
			pedestrian	**rider**	**average**	**pedestrian**	**rider**	**average**
w/o $\mathcal{L}_{hfc}$	12.25	4.69	78.2	65.7	71.9	38.6	59.1	48.8
smaller training epochs	14.86	4.21	78.8	66.8	72.8	38.5	55.9	47.2
smaller input size	27.21	4.73	77.0	66.1	71.5	41.7	56.9	49.3
Ours	12.17	4.22	79.0	66.6	72.8	37.9	54.0	45.9

FID quantifies the distributional distance between synthetic and real-world nighttime images. NIQE assesses the quality of real-world nighttime images after enhancement. mAP and LAMR(all) measure detection accuracy.

(a) w/o Illumination Brightening (b) w/o Denoising (c) w/o ID (d) w/o PSA

(e) w/o DA (f) Ours (g) Ground Truth

Fig. 6. Ablation study of individual components in the proposed method. (a) Illumination Brightening. (b) Illumination Difference-Aware Denoising Network. (c) Illumination Difference Guide. (d) Psuedo-Supervised Attention. (e) Degradation Alignment. (f) Ours. (g) Ground Truth. The pink bounding box represents the pedestrian class. Zoom in for better view.

degradation alignment, we remove the domain classifier, gradient reversal layer, and real-world nighttime image input from the denoising network. As shown in Fig. 6(e), although the result without degradation alignment appears visually similar to ours, the quantitative results in the third row of Table VII reveal that degradation alignment enhances noise suppression in real-world nighttime images and preserves more informative features in the denoised output, ultimately improving detection performance.

D. Analysis of the Effect of Synthetic Image Quality on Performance of Denoising Network

To better understand how the quality of synthetic data impacts the performance of the denoising network, we modify the training configuration of the generative model. Specifically, we introduce the following variations: (1) the removal of

TABLE IX
COMPARISON OF RESULTS ON NIGHTOWLS-1K.

Category	Method	LAMR(%) ↓			mAP(%) ↑		
		pedestrian	rider	average	pedestrian	rider	average
Generalization	YOLOv8s [64]	69.3	57.9	63.6	57.4	43.5	50.4
Enhancement Based (with YOLOv8s)	ZeroDCE [17]	72.3	66.3	69.3	53.2	29.2	41.2
	ZeroDCE* [17]	77.4	94.2	85.8	47.9	5.6	26.7
	EnlightenGAN [16]	74.7	60.5	67.6	50.3	39.5	44.9
	EnlightenGAN* [16]	68.0	55.5	61.7	58.2	48.2	53.2
	RetinexNet [14]	78.8	77.3	78.0	46.2	26.5	36.3
	JED [29]	69.7	66.3	68.0	56.8	30.9	43.8
	Retinexformer [15]	73.8	69.5	71.6	53.2	30.4	41.8
	LIME [28]	73.4	79.1	76.2	53.1	19.5	36.3
	SNR [65]	71.9	69.0	70.4	54.6	27.1	40.8
	KinD [37]	71.7	59.5	65.6	53.3	43.3	48.3
	KinD++ [66]	75.9	67.7	71.8	47.8	33.8	40.8
	GLARE [67]	68.0	65.9	66.9	58.9	34.5	46.7
	CIDNet [68]	73.0	65.9	69.4	53.0	32.2	42.6
Unsupervised Domain Adaptation	EPM [12]	80.7	55.8	68.2	36.9	29.1	33.0
	SIGMA++ [13]	78.9	56.4	67.6	39.9	21.5	30.7
	ALDI++ [69]	70.0	25.9	47.9	51.7	69.9	60.8
	Simple SFOD [70]	99.5	100.0	99.7	1.8	0.0	0.9
	JED† [29]	70.9	43.3	57.1	56.0	51.0	53.5
	GLARE† [67]	71.2	46.1	58.6	56.2	51.5	53.8
	Ours	59.3	38.6	48.9	68.6	64.0	66.3

All models were trained on the ECP dataset and directly tested on NightOwls-1k. (Key: Best, Second Best, Third Best)

the proposed high-frequency consistency loss ($\mathcal{L}_{hfc}$), (2) a reduction in the number of training epochs from 100 to 80, and (3) a decrease in the input image size from 360×360 to 256×256.

To quantify the alignment between synthetic and real-world nighttime image distributions, we employ the Fréchet Inception Distance (FID) [71]. The denoising task aims to remove noise contamination and restore textural details while maintaining semantic integrity. Therefore, its performance is evaluated using both low-level (image quality) and high-level (detection accuracy) metrics. Given the absence of a noise-free nighttime reference image, we use the reference-free image quality assessment metric NIQE [72]. Detection performance is evaluated using mAP and LAMR(all).

The quantitative results in Table VIII reveal that a large domain gap between synthetic and real-world nighttime images can impair the denoising network's robustness on real-world data, leading to lower-quality enhanced images and decreased detection performance. Notably, the synthetic nighttime images generated by our method most closely resemble real-world nighttime images. As a result, our approach achieves the most balanced performance across the three metrics. This suggests that our method not only recovers more visually appealing textural details but also preserves semantic integrity.

TABLE X
SENSITIVITY ANALYSIS OF λ_{ps} AND λ_{dc}.

λ_{ps}	λ_{dc}	LAMR(all) ↓	mAP ↑
1.5×10^2	1	45.3	72.6
0.5×10^2	1	45.6	72.7
1×10^2	1.5	46.2	73.5
1×10^2	0.5	46.7	72.4
1×10^2	1	45.9	72.8

The scaled weights are indicated in blue.

E. Sensitivity Analysis of Loss Function Weights

In addition to the commonly used loss functions in image restoration, two novel loss functions are introduced to optimize the denoising network. A sensitivity analysis is conducted to assess model robustness by scaling the weights of pseudo-supervision loss (λ_{ps}) and domain classification loss (λ_{dc}), while keeping λ_{rec}, λ_{col}, and λ_{per} fixed. The results in Table X show that scaling the weights by 50% has minimal impact on performance, suggesting the method's robustness and eliminating the need for meticulous hyperparameter tuning.

F. Applications

To further validate the effectiveness and robustness of the proposed method across a broader range of real-world nighttime scenarios, two additional test sets are used, denoted as NightOwls-1k and WHUCampus. NightOwls-1k consists of 1,000 images randomly selected from the NightOwls validation set [41], which includes four categories: pedestrians, bicycle drivers, motorbike drivers, and ignored areas. To match the class settings of the ECP dataset, the bicycle driver and motorbike driver categories were merged into a single "rider" class. Additionally, we collected nighttime images using an iPhone XS smartphone on the campus of Wuhan University, forming the WHUCampus dataset, which contains 203 valid images. The size of each image is 1920×1080. We manually labeled pedestrians and riders in these images to create the ground truth. These test images are characterized by significant motion blur and image noise, presenting considerable challenges for processing.

Performance on NightOwls-1k. The quantitative results are presented in Table IX, while the qualitative comparisons of the top five performing methods are illustrated in Fig. 7.

Fig. 7. Comparison of test results for the top five performing methods on NightOwls-1k test set. The pink bounding box represents the pedestrian class and the green bounding box represents the rider class.

Notably, all enhancement methods, except for the retrained EnlightenGAN, degrade detection performance, underscoring the importance of addressing domain gaps. ALDI++ demonstrates strong performance in detecting riders, primarily due to generating numerous prediction bounding boxes. This approach reduces the LAMR at the cost of significantly increasing false positives. In contrast, our method achieves a higher mAP while maintaining a comparable LAMR to ALDI++. This result suggests that our enhancement techniques effectively recover sharper details and textures, leading to improved detection precision.

Performance on WHUCampus. The quantitative results are presented in Table XI. Our method, along with GLARE† and JED†, outperforms other approaches, emphasizing the significance of distribution alignment. The qualitative results for the top five performing methods are displayed in Fig. 9, where our method demonstrates superior accuracy in detecting and classifying riders. This suggests that our enhanced results exhibit more informative image features, which are beneficial for the detector. The top performance in both metrics underscores the robustness of our proposed framework in challenging nighttime environments.

G. t-SNE Visualization

We perform t-distributed stochastic neighbor embedding (t-SNE) [73], a high-dimensional data visualization technique, to visualize the data distributions of day, night, and synthetic night images. We first use a pretrained ResNet-101 [74] to extract features from images, and then input the high-dimensional features into t-SNE to reduce dimensions so that it can be plotted in a scatter map. A visualization of the ECP and the KAIST datasets is displayed in Fig. 8. Every data point corresponds to an image. The results indicate that there exists a huge domain gap between daytime and nighttime distributions. It is difficult to distinguish between the representations of synthetic and real-world nighttime images, which are highly confused. It illustrates that the distribution alignment can successfully transfer daytime images into the

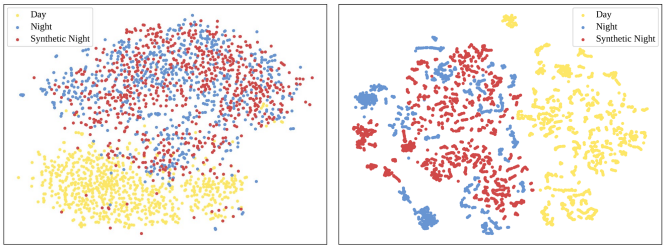
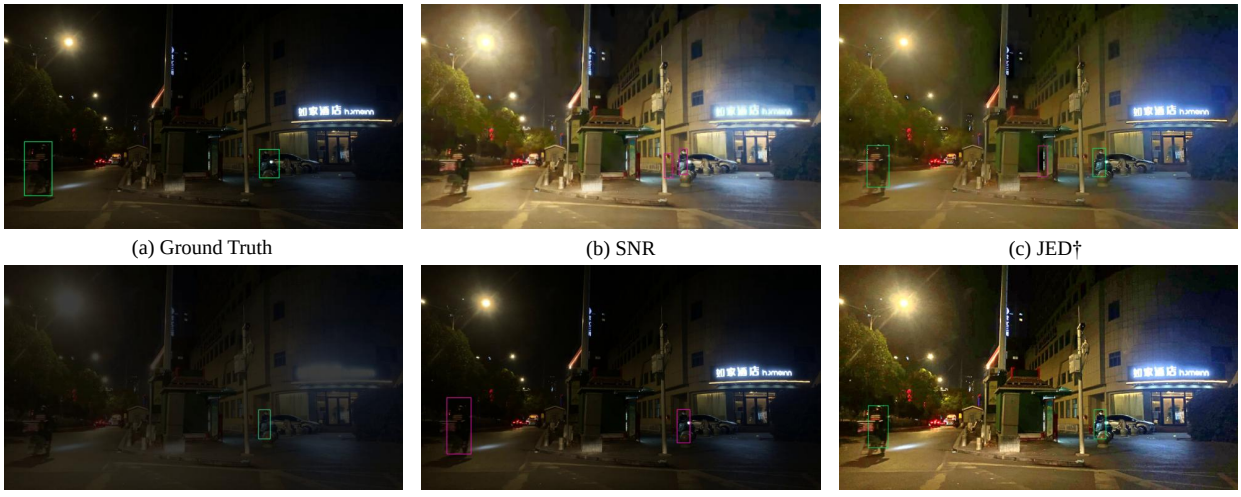


Fig. 8. t-SNE [73] visualization of images' features extracted by pretrained ResNet-101 [74] on (a) the ECP dataset and (b) the KAIST dataset. Yellow, blue, and red dots represent the features of daytime, nighttime, and synthetic nighttime images, respectively.

TABLE XI
COMPARISON OF RESULTS ON WHUCAMPUS.

Category	Method	LAMR(%) ↓			mAP(%) ↑		
		pedestrian	rider	average	pedestrian	rider	average
Generalization	YOLOv8s [64]	62.6	82.9	72.7	62.0	46.2	54.1
Enhancement Based (with YOLOv8s)	ZeroDCE [17]	75.2	87.3	81.2	50.0	45.8	47.9
	ZeroDCE* [17]	73.1	85.6	79.3	51.0	42.6	46.8
	EnlightenGAN [16]	72.5	78.4	75.4	51.1	46.3	48.7
	EnlightenGAN* [16]	73.2	75.7	74.4	54.0	55.3	54.6
	RetinexNet [14]	92.0	97.0	94.5	24.1	37.1	30.6
	JED [29]	74.0	78.5	76.2	52.7	47.3	50.0
	Retinexformer [15]	80.2	88.0	84.1	45.4	46.7	46.0
	LIME [28]	75.8	85.6	80.7	46.5	43.8	45.1
	SNR [65]	61.7	71.9	66.8	63.6	54.9	59.2
	KinD [37]	64.9	81.8	73.3	60.1	45.3	52.7
	KinD++ [66]	75.3	82.6	78.9	48.8	40.8	44.8
	GLARE [67]	58.9	76.2	67.5	64.7	56.6	60.6
	CIDNet [68]	75.7	81.8	78.7	48.9	50.9	49.9
Unsupervised Domain Adaptation	EPM [12]	78.6	78.5	78.5	36.0	26.4	31.2
	SIGMA++ [13]	79.7	71.9	75.8	37.3	38.7	38.0
	ALDI++ [69]	55.7	62.7	59.2	63.5	55.2	59.3
	Simple SFOD [70]	99.6	100.0	99.8	2.5	0.0	1.2
	JED† [29]	56.8	60.5	58.6	67.6	62.8	65.2
	GLARE† [67]	51.3	59.4	55.3	72.7	65.7	69.2
	Ours	49.4	55.7	52.5	74.7	68.3	71.5

All models were trained on the ECP dataset and directly tested on WHUCampus. (Key: Best, Second Best, Third Best)



(a) Ground Truth (b) SNR (c) JED†

(d) GLARE† (e) ALDI++ (f) Ours

Fig. 9. Comparison of test results for the top five performing methods on WHUCampus test data. The pink bounding box represents the pedestrian class and the green bounding box represents the rider class.

nighttime domain, producing sufficiently realistic synthetic nighttime images to fool the feature extractor.

V. CONCLUSION

In this paper, we present an unsupervised domain adaptation framework for nighttime VRU detection, which frames the task as domain distribution alignment and nighttime image enhancement. We introduce a high-frequency consistency loss in a generative model to align the distributions of daytime and nighttime images, generating synthetic nighttime images. The degradations in these images are then decoupled and enhanced separately. First, the illumination is brightened, followed by denoising using our proposed illumination difference-aware denoising network. To supervise this network, we design an exchange-recombination strategy to generate pseudo-ground truth, guiding the denoising process, and implement pixel-level supervision through pseudo-supervised attention. Additionally, we introduce degradation alignment in the denoising network, encouraging the model to learn robust degradation features, thereby enhancing its performance on real-world nighttime images. Extensive qualitative and quantitative experimental results demonstrate the effectiveness of our approach, highlighting its superiority and generalization capability.

Although the proposed framework does not require nighttime annotations, it assumes that daytime images do not experience significant degradation. This assumption may limit the

method's applicability in scenarios where daytime images are severely degraded or unavailable. Future research will explore strategies to mitigate this dependency on daytime images. Potential approaches include zero-shot image enhancement techniques or translating degraded nighttime images into the daytime domain, thereby enabling the use of existing detection models trained on daytime data.

REFERENCES

[1] European Commission. Directorate General for Mobility and Transport., *Supporting Study on Activities 3.2, 3.3 and 3.4 of the New Working Programme of the ITS Directive: Final Report.* LU: Publications Office, 2021.

[2] X. Lv, Z. Xiao, J. Fang, Q. Li, F. Lei, and G. Sun, "On safety design of vehicle for protection of vulnerable road users: A review," *Thin-Walled Structures*, vol. 182, p. 109990, 2023.

[3] G. Yuan, T. Ye, H. Fu, L. Wang, and Z. Wang, "Clustering based detection of small target pedestrians for smart cities," *Sustainable Energy Technologies and Assessments*, vol. 52, p. 102300, Aug. 2022.

[4] A. Ben Khalifa, I. Alouani, M. A. Mahjoub, and A. Rivenq, "A novel multi-view pedestrian detection database for collaborative Intelligent Transportation Systems," *Future Generation Computer Systems*, vol. 113, pp. 506–527, Dec. 2020.

[5] N. Kirillova, H. Possegger, and H. Bischof, "A Visual Surveillance System to Observe Realistic Road User Behavior for Improved Pedestrian and Cyclist Safety at Crossroads," in *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. Madrid, Spain: IEEE, Nov. 2022, pp. 1–8.

[6] J. Li, Y. Bi, S. Wang, and Q. Li, "CFRLA-Net: A context-aware feature representation learning anchor-free network for pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4948–4961, 2023.

[7] K. Kumar and R. K. Mishra, "A diagonally oriented novel feature extractor for pedestrian detection and its efficient hardware implementation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2035–2042, 2021.

[8] A. H. Khan, M. S. Nawaz, and A. Dengel, "Localized Semantic Feature Mixers for Efficient Pedestrian Detection in Autonomous Driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5476–5485.

[9] C. Jiang, H. Xu, W. Zhang, X. Liang, and Z. Li, "SP-NAS: Serial-to-parallel backbone search for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 863–11 872.

[10] Z. Lin, W. Pei, F. Chen, D. Zhang, and G. Lu, "Pedestrian detection by exemplar-guided contrastive learning," *IEEE transactions on image processing*, vol. 32, pp. 2003–2016, 2022.

[11] A. Lengyel, S. Garg, M. Milford, and J. C. Van Gemert, "Zero-Shot Day-Night Domain Adaptation with a Physics Prior," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 4379–4389.

[12] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin, and M.-H. Yang, "Every Pixel Matters: Center-Aware Feature Alignment for Domain Adaptive Object Detector," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, vol. 12354, pp. 733–748.

[13] W. Li, X. Liu, and Y. Yuan, "SIGMA++: Improved Semantic-complete Graph Matching for Domain Adaptive Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2023.

[14] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex Decomposition for Low-Light Enhancement," Aug. 2018.

[15] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage Retinex-based Transformer for Low-light Image Enhancement," Aug. 2023.

[16] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep Light Enhancement Without Paired Supervision," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 30, p. 10, 2021.

[17] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 1777–1786.

[18] W. Wang, Z. Xu, H. Huang, and J. Liu, "Self-aligned concave curve: Illumination enhancement for unsupervised adaptation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2617–2626.

[19] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, Jun. 2015, pp. 1037–1045.

[20] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "Eurocity persons: A novel benchmark for person detection in traffic scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1844–1861, 2019.

[21] H. Wang, S. Liao, and L. Shao, "AFAN: Augmented Feature Alignment Network for Cross-Domain Object Detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 4046–4056, 2021.

[22] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, and M.-H. Yang, "Progressive Domain Adaptation for Object Detection."

[23] L. Fu, H. Yu, F. Juefei-Xu, J. Li, Q. Guo, and S. Wang, "Let There Be Light: Improved Traffic Surveillance via Detail Preserving Night-to-Day Transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8217–8226, Dec. 2022.

[24] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-GANs: Edge-Aware Generative Adversarial Networks for Cross-Modality MR Image Synthesis," *IEEE Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1750–1762, Jul. 2019.

[25] Y. Wang, Q. Chen, and B. Zhang, "Image enhancement based on equal area dualistic sub-image histogram equalization method," *IEEE transactions on Consumer Electronics*, vol. 45, no. 1, pp. 68–75, 1999.

[26] X. Dong, Y. Pang, and J. Wen, "Fast efficient algorithm for enhancement of low lighting video," in *ACM SIGGRAPH 2010 Posters*, 2010, pp. 1–1.

[27] G. Kim and J. Kwon, "Deep Illumination-Aware Dehazing With Low-Light and Detail Enhancement," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2494–2508, Mar. 2022.

[28] X. Guo, Y. Li, and H. Ling, "LIME: Low-Light Image Enhancement via Illumination Map Estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 982–993, Feb. 2017.

[29] X. Ren, M. Li, W.-H. Cheng, and J. Liu, "Joint Enhancement and Denoising Method via Sequential Decomposition," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. Florence: IEEE, May 2018, pp. 1–5.

[30] S. Hao, X. Han, Y. Guo, X. Xu, and M. Wang, "Low-Light Image Enhancement With Semi-Decoupled Decomposition," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3025–3038, Dec. 2020.

[31] X. Guo and Q. Hu, "Low-light Image Enhancement via Breaking Down the Darkness," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 48–66, Jan. 2023.

[32] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired Unrolling with Cooperative Prior Architecture Search for Low-light Image Enhancement," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 10 556–10 565.

[33] Q. Jiang, Y. Mao, R. Cong, W. Ren, C. Huang, and F. Shao, "Unsupervised Decomposition and Correction Network for Low-Light Image Enhancement," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–16, 2022.

[34] Z. Chen, Y. Jiang, D. Liu, and Z. Wang, "CERL: A Unified Optimization Framework for Light Enhancement With Realistic Noise," *IEEE Transactions on Image Processing*, vol. 31, pp. 4162–4172, 2022.

[35] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward Fast, Flexible, and Robust Low-Light Image Enhancement," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 5627–5636.

[36] C. Li, C. Guo, and C. L. Chen, "Learning to Enhance Low-Light Image via Zero-Reference Deep Curve Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[37] Y. Zhang, J. Zhang, and X. Guo, "Kindling the Darkness: A Practical Low-light Image Enhancer," in *Proceedings of the 27th ACM International Conference on Multimedia*. Nice France: ACM, Oct. 2019, pp. 1632–1640.

[38] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3213–3221.

[39] Xiaofei Li, F. Flohr, Yue Yang, Hui Xiong, M. Braun, S. Pan, Keqiang Li, and D. M. Gavrila, "A new benchmark for vision-based cyclist detection," in *2016 IEEE Intelligent Vehicles Symposium (IV)*. Gotenburg, Sweden: IEEE, Jun. 2016, pp. 1028–1033.

[40] X. Dai, J. Hu, C. Luo, H. Zerfa, H. Zhang, and Y. Duan, "NIRPed: A Novel Benchmark for Nighttime Pedestrian and Its Distance Joint Detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 6932–6942, Jul. 2023.

[41] L. Neumann, M. Karg, S. Zhang, C. Scharfenberger, E. Piegert, S. Mistr, O. Prokofyeva, R. Thiel, A. Vedaldi, and A. Zisserman, "Nightowls: A pedestrians at night dataset," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part I 14*. Springer, 2019, pp. 691–705.

[42] N. Chen, J. Xie, J. Nie, J. Cao, Z. Shao, and Y. Pang, "Attentive Alignment Network for Multispectral Pedestrian Detection," in *Proceedings of the 31st ACM International Conference on Multimedia*. Ottawa ON Canada: ACM, Oct. 2023, pp. 3787–3795.

[43] G. Golcarenarenji, I. Martinez-Alpiste, Q. Wang, and J. M. Alcaraz-Calero, "Illumination-aware image fusion for around-the-clock human detection in adverse environments from Unmanned Aerial Vehicle," *Expert Systems with Applications*, vol. 204, p. 117413, 2022.

[44] G. Li, S. Zhang, and J. Yang, "Nighttime Pedestrian Detection Based on Feature Attention and Transformation," in *2020 25th International Conference on Pattern Recognition (ICPR)*. Milan, Italy: IEEE, Jan. 2021, pp. 9180–9187.

[45] H. Yao, Y. Zhang, H. Jian, L. Zhang, and R. Cheng, "Nighttime pedestrian detection based on Fore-Background contrast learning," *Knowledge-Based Systems*, vol. 275, p. 110719, Sep. 2023.

[46] J. Liu, D. Xu, W. Yang, M. Fan, and H. Huang, "Benchmarking Low-Light Image Enhancement and Beyond," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1153–1184, Apr. 2021.

[47] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 1792–1800, Jun. 2022.

[48] Q. Qin, K. Chang, M. Huang, and G. Li, "DENet: Detection-driven Enhancement Network for Object Detection Under Adverse Weather Conditions," in *Computer Vision – ACCV 2022*, L. Wang, J. Gall, T.-J. Chin, I. Sato, and R. Chellappa, Eds. Cham: Springer Nature Switzerland, 2023, vol. 13843, pp. 491–507.

[49] Z. Cui, G.-J. Qi, L. Gu, S. You, Z. Zhang, and T. Harada, "Multitask AET with Orthogonal Tangent Regularity for Dark Object Detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 2533–2542.

[50] W. Wang, X. Wang, W. Yang, and J. Liu, "Unsupervised Face Detection in the Dark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1250–1266, Jan. 2023.

[51] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Aug. 2021.

[53] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform," Apr. 2018.

[54] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," *International conference on machine learning. PMLR,*, 2015.

[55] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5301–5310.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[57] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," *arXiv preprint arXiv:1808.04818*, 2018.

[58] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[60] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[61] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

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

[63] G. Jocher, "YOLOv5 by ultralytics," 2020.

[64] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023.

[65] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "SNR-Aware Low-light Image Enhancement," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 17 693–17 703.

[66] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond Brightening Low-light Images," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1013–1037, Apr. 2021.

[67] H. Zhou, W. Dong, X. Liu, S. Liu, X. Min, G. Zhai, and J. Chen, "GLARE: Low Light Image Enhancement via Generative Latent Feature based Codebook Retrieval," Jul. 2024.

[68] Q. Yan, Y. Feng, C. Zhang, P. Wang, P. Wu, W. Dong, J. Sun, and Y. Zhang, "You Only Need One Color Space: An Efficient Network for Low-light Image Enhancement," Jun. 2024.

[69] J. Kay, T. Haucke, S. Stathatos, S. Deng, E. Young, P. Perona, S. Beery, and G. Van Horn, "Align and Distill: Unifying and Improving Domain Adaptive Object Detection," Aug. 2024.

[70] Y. Hao, F. Forest, and O. Fink, "Simplifying Source-Free Domain Adaptation for Object Detection: Effective Self-training Strategies and Performance Insights," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, vol. 15112, pp. 196–213.

[71] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[72] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[73] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

Yuankun Wang received the B.S. degree in remote sensing science and technology from China University of Geosciences, Wuhan, China, in 2022. She is currently working toward the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan China. Her current research interests include computer vision and deep learning. She mainly works in the area of image enhancement and domain adaptation-based object detection.

Zhenfeng Shao received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2004. Since 2009, he has been a Full Professor with LIESMARS, Wuhan University. He has authored or coauthored over 50 peer-reviewed articles in international journals. His research interests include high-resolution image processing, pattern recognition, and urban remote sensing applications. Dr. Shao was a recipient of the Talbert Abrams Award for the Best Paper in Image matching from the American Society for Photogrammetry and Remote Sensing in 2014 and the New Century Excellent Talents in University from the Ministry of Education of China in 2012. He has served as an Associate Editor of the Photogrammetric Engineering and Remote Sensing (PE and RS) specializing in smart cities, photogrammetry and change detection since 2019.

Jiaming Wang received his Ph.D. degree in photogrammetry and remote sensing from Wuhan University in 2022. He is currently a teacher in School of Computer Science and Engineering, Wuhan Institute of Technology. His research field includes image/video processing, computer vision.

Yu Wang received the master's degree from Wuhan Institute of Technology, Wuhan, China, in 2021. He is currently pursuing a Ph.D. degree under the supervision of Prof. Zhenfeng Shao with the State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan. His research field includes image/video processing and computer vision.

Yulin Ding received the B.S. degree in geographic information science from Northwest University, Xi'an, China, in 2020. She is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan China. Her research interests include impervious surface extraction based on SAR imagery and remote sensing image semantic segmentation.

Gui Cheng received the B.S degree in geographic information science from Chang'an University, Xi'an, China in 2019. He is currently working toward the Ph.D. degree in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. His research interests include image processing and computer vision.