


Article

# Zero-Shot Day–Night Domain Adaptation for Face Detection Based on DAI-CLIP-Dino

Huadong Sun \*, Yinghui Liu, Ziyang Chen and Pengyi Zhang 

School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China; lyh@s.hrbcu.edu.cn (Y.L.); chenziyang@s.hrbcu.edu.cn (Z.C.); zpy@s.hrbcu.edu.cn (P.Z.)

\* Correspondence: sunhuadong1209@163.com

**Abstract:** Two challenges in computer vision (CV) related to face detection are the difficulty of acquisition in the target domain and the degradation of image quality. Especially in low-light situations, the poor visibility of images is difficult to label, which results in detectors trained under well-lit conditions exhibiting reduced performance in low-light environments. Conventional works image enhancement and object detection techniques are unable to resolve the inherent difficulties in collecting and labeling low-light images. The Dark-Illuminated Network with Contrastive Language–Image Pretraining (CLIP) and Self-Supervised Vision Transformer (Dino), abbreviated as DAI-CLIP-Dino is proposed to address the degradation of object detection performance in low-light environments and achieve zero-shot day–night domain adaptation. Specifically, an advanced reflectance representation learning module (which leverages Retinex decomposition to extract reflectance and illumination features from both low-light and well-lit images) and an interchange–redecomposition coherence process (which performs a second decomposition on reconstructed images after the exchange to generate a second round of reflectance and illumination predictions while validating their consistency using redecomposition consistency loss) are employed to achieve illumination invariance and enhance model performance. CLIP (ViT-based image encoder part) and Dino have been integrated for feature extraction, improving performance under extreme lighting conditions and enhancing its generalization capability. Our model achieves a mean average precision (mAP) of 29.6% for face detection on the DARK FACE dataset, outperforming other models in zero-shot domain adaptation for face detection.



Academic Editor: Beiwen Li

Received: 25 October 2024

Revised: 28 November 2024

Accepted: 18 December 2024

Published: 1 January 2025

**Citation:** Sun, H.; Liu, Y.; Chen, Z.; Zhang, P. Zero-Shot Day–Night Domain Adaptation for Face Detection Based on DAI-CLIP-Dino. *Electronics* **2025**, *14*, 143. <https://doi.org/10.3390/electronics14010143>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** face detection; zero-shot day–night domain adaptation; DAI-CLIP-Dino

## 1. Introduction

In practice, the drastic degradation of the model’s test performance on non-distributed data has led to severe limitations in its application [1]. Due to insufficient light and uneven exposure in low-light environments, low-light images suffer from high image noise, low signal-to-noise ratios, low contrast, and color distortion. Furthermore, in low-light environments, image blurriness, insufficient dynamic range, and effects of reflections and light patches may result in an insignificant difference between object and background, thereby making detection ineffective, which greatly hampers the performance of models. This performance decline could become a limiting factor in any critical application that relies on visual information. Two common strategies are proposed to address this challenge, including image enhancement methods to improve image visibility under low-light conditions [2,3]; fine-tuning detectors are originally trained on well-illuminated images [4,5].

Image enhancement methods are frequently effective, while they rely on a huge number of low-light images collected in the real world. Current methods require training with compared low-light and well-lit pictures [6], and they also require an increased utilization of low-light visual data for the detection of dark objects [7]. Compared to datasets with good lighting, low-light image datasets are difficult to collect [8]. Due to low visibility, annotating bounding boxes is extremely challenging, and this hinders the improvement of low-light picture enhancement and object detection. Common datasets with good lighting include WIDER FACE [9], i.e., 32,203 images and 393,703 labeled faces, whereas the DARK FACE dataset, under low-light conditions, contains only 10,000 images and 81,560 faces.

To successfully handle the difficulty of object identification in low-light conditions, we present a zero-shot day–night domain adaptation strategy. Operating within this model, the object detector is trained in the brightly illuminated source domain and assessed in the dimly lit target domain, without the input of any actual images. This method primarily uses the disparity in illumination between the source and target domains to investigate the influence of lighting variations on detection accuracy.

The proposed strategy is shown in Figure 1, where the zero-shot day–night domain adaptation strategy adopts the Retinex [10] image decomposition method, which posits that an image can be decomposed into reflectance and illumination components [11]. Reflectance represents the illumination-invariant information critical for low-light object detection, while illumination affects the image’s visibility. We integrate this method into established object detection frameworks, such as DSFD [12], by incorporating a module to learn reflectance representation as a decoder. The redecomposition cohering loss enhances the Retinex-based image decomposition process as a key technique. By introducing a two-round decomposition process, where the reflectance is interchanged and redecomposed to generate new reflectance, this approach ensures the coherence between the two rounds of reflectance, thereby improving the stability and accuracy of the reflectance representation. The main contributions of this paper are summarized as follows:

- ZSDA is engineered to effectively decode reflectance-based illumination-invariant data from both naturally well-lit and synthetically produced low-light images. The pre-trained RetinexNet [10] network is utilized to further enhance this module, with specific illumination invariance enhancement strategies to boost its performance.
- The exchange–decomposition coherence process [13] is proposed to improve the quality of image decomposition based on the Retinex theory. This process enhances reflectance consistency by introducing recomposition coherence loss during two decomposition stages, thereby improving the stability and accuracy of image reconstruction.
- ZSDA allows the model to be trained solely on well-lit source domain images and to perform precise evaluations in completely image-less low-light target domains, which enhances adaptability and generalization of model under extreme lighting variations. Additionally, we further enhance the model’s capabilities by merging the image encoder of CLIP [14] (Contrastive Language Image Pre-training) and the technology of Dino [15] (Self-Distillation with No Labels).
- The generalization ability of the CLIP (ViT-based image encoder part) and the self-supervised learning characteristics of Dino are used to jointly improve the model’s capacity to capture details in low-light environments and understand complex scenes; this enables the model to more accurately recognize and process images in dim lighting conditions, improving the accuracy and reliability of object detection.

The remainder of paper is organized as follows: Section 2 provides a comprehensive overview of the related work in the field, focusing on key areas such as object detection, low-light images, and zero-shot domain adaptation. The methodology for zero-shot day–

night domain adaptation for face detection is provided in Section 3, detailing the innovative approaches and techniques employed. The analysis of the experiments and results is described in Section 4.

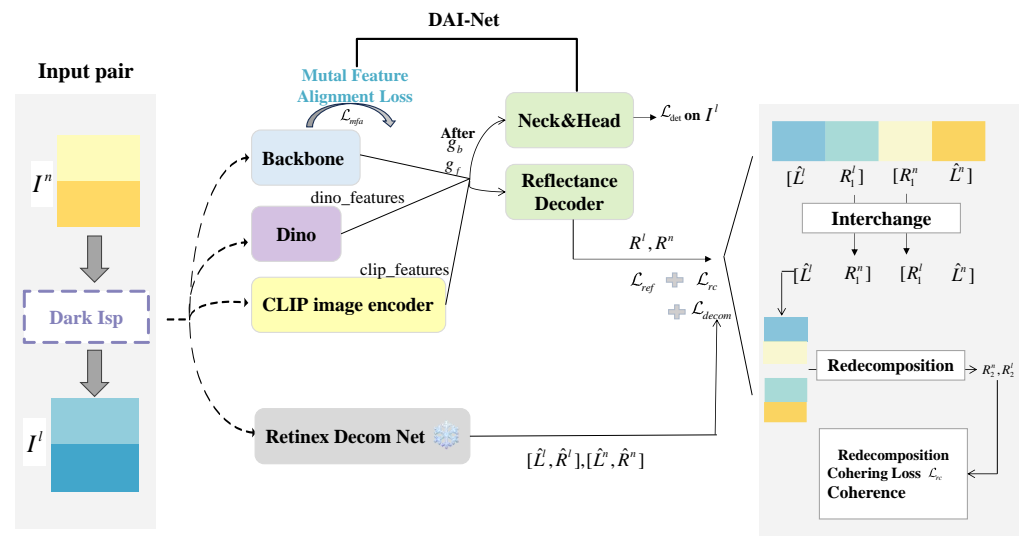


Figure 1. Framework Outline of the Methodology.

## 2. Related Work

### 2.1. Object Detection

Detectors are primarily categorized into two types: single-stage and two-stage. Single-stage detectors, such as SSD [16], YOLO [17], and FCOS [18], are designed to simplify the detection process, directly predicting the bounding boxes and categories of objects in one pass, which enhances processing speed. In contrast, two-stage detectors like Faster R-CNN [19] and R-FCN [20] focus more on accuracy. The initial step involves generating candidate areas; subsequently, classification and bounding box regression are performed on these regions to achieve more precise detection results. Face detectors [12,21–23] mostly use the single-stage approach, which efficiently outputs both bounding boxes and class scores simultaneously. Additionally, current research continuously explores more efficient model architectures [24], improved anchor sampling techniques [25], and feature enhancement technologies [26] to boost detection performance. Recently, transformer-based detection models such as DETR [27] and its derivatives have significantly enhanced overall object detection capabilities through global context optimization and an end-to-end training architecture, demonstrating substantial performance improvements.

Object detection techniques have made great strides due to the widespread use of datasets such as COCO [28] and Open Images [29], as well as face detection-specific datasets such as WIDER FACE. These datasets provide a wealth of material for researchers to train, test and refine detection models.

### 2.2. Processing of Low-Light Images

The advancement of deep learning technologies has fostered breakthroughs in low-light image-enhancing technology, which primarily encompass two approaches. One approach involves enhancement via reflectance, exemplified by [2], who modeled both reflectance and illuminance degradation using a transformer structure. The other approach involves enhancing images by adjusting lighting through reconstruction, such as [10], who utilized a Retinex decomposition model for light enhancement and denoising, addressing noise based on reflectance and illuminance mapping [30], and optimizing the process

based on the Retinex model, employing a cooperative dual-layer search strategy to find the desired network structure.

Common low-light image target detection is mainly classified into three categories: the first category of methods is low-light enhancement methods, which are used to generate bright images to achieve the purpose of visual enhancement; the second category of methods is the training process of the detection model, i.e., the integration of image enhancement techniques to be used to improve the performance of object detection; the last category of methods comprises low-light detector learning strategies, including multi-model merging [31], multi-task auto-coding [7], and unsupervised domain-adaptive framework [5], which improve the direct detection ability of the detector in low-light conditions. This enables them to achieve the improvement of the detection model's direct detection ability in low-light conditions and make the detector performance more robust in low-light environments. Within these methodologies, strategies tailored to specific domains have garnered significant attention. As an example, ref. [32] presents a detection framework specifically designed for nighttime conditions. By incorporating vehicle-specific features such as headlights and taillights, in combination with a multi-level fusion network and hierarchical labeling, this approach addresses the limitations of conventional enhancement techniques and domain-centric detection methods. Its innovative application of highlight information offers a robust solution for vehicle detection under low-light conditions, thereby expanding the scope and effectiveness of this class of methods.

### 2.3. Zero-Shot Domain Adaptation

Domain-invariant representation aims to transfer information from the source domain to the destination domain, particularly when the target domain has a scarcity or absence of annotated data. The core is constructing domain invariant representations and adopting various strategies to reduce the feature distribution differences between domains.

Unsupervised domain adaptation (UDA) primarily includes methods such as adversarial learning [11], self-training [33], entropy minimization [34], and generative-based adaptation [35]. These methods achieve effective knowledge transfer by reducing domain discrepancies at the input, feature, or output levels. However, some approaches designed to mitigate negative migration effects still rely on data from the target domain. In contrast, zero-shot domain adaptation (ZSDA) presents more complex scenarios by requiring solutions to domain distribution discrepancies in the complete absence of target domain data. This approach highlights key challenges in domain adaptation research and provides innovative solutions. In particular, the introduction of physical prior processing underscores the significance of novel methodologies in zero-sample settings without target domain images. These advances not only deepen our understanding of cross-domain knowledge transfer but also offer new pathways to tackle adaptation challenges in practical applications. Furthermore, generative models like coupled generative adversarial networks (CoGANs) [36] and variational autoencoders (VAEs) [37] have been used in ZSDA to reconstruct target domain samples, but these approaches come with substantial computational costs.

Research has focused on the study of learning in dark environments, specifically object segmentation and detection. This research falls under the umbrella of domain transfer learning, which is further separated into two categories: domain adaptation (DA) and domain generalization (DG). Domain adaptation (DA) strategies include training models using well-illuminated data from a source domain in order to adapt to low-light data from a target domain. Common techniques involve the creation of low-light pictures through synthesis [7,38], aligning distributions in well-lit and low-light domains through self-supervised learning [5], merging components from both domains [11,34], and other multi-stage strategies [39–41]. DG, unlike DA, generalizes to unknown domains without

prior knowledge of the target domain [42]. Zero-shot day–night domain adaptation is a method that addresses low-light settings. It focuses on a particular example of domain adaptation (DA) when genuine low-light data are not available, but the target domain is recognized as a low-light environment.

The DAI-Net [13] framework incorporates Retinex theory to achieve robust domain-invariant representation learning. Advanced feature extraction modules, such as the CLIP [14] image encoder and DINO [15], are integrated into the framework. The CLIP image encoder leverages large-scale multimodal pre-training to provide powerful global semantic understanding, while DINO extracts fine-grained, self-supervised features, enabling the model to maintain stable feature representations under varying illumination conditions. This integration enhances the framework’s ability to bridge the illumination gap between well-lit and low-light environments, addressing key challenges in ZSDA tasks by ensuring effective feature alignment and strong generalization without relying on real target domain data.

Despite significant advancements in object detection, low-light image processing, and domain adaptation, existing methods face several limitations:

- **Global and local feature extraction imbalance:** models often fail to balance global and local feature extraction, resulting in incomplete feature representations, particularly in low-light environments.
- **Feature degradation in low-light conditions:** severe lighting variations lead to degraded feature quality, while current methods lack robust illumination-invariant feature modeling.
- **Weak domain-invariant feature learning:** insufficient mechanisms for domain-invariant representation learning hinder performance in zero-shot domain adaptation tasks.
- **Limited generalization:** adapting to complex scenarios, such as day–night transitions, remains a challenge due to inadequate integration of global and local features.

To address these challenges, this work builds upon the DAI-Net framework, integrating Retinex theory with advanced feature extraction capabilities provided by the CLIP image encoder and DINO. This approach effectively resolves the aforementioned issues, achieving robust feature alignment, illumination invariance, and superior performance in zero-shot domain adaptation tasks.

### 3. Method

The task at hand involves training an object detector using a well-illuminated picture and then applying it to a low-light image. This is known as a zero-shot day–night domain adaptation issue, where the objective is to adapt the detector from a well-lit source domain to a low-light target domain. The primary difficulty lies in the variation in lighting conditions, which leads to a decline in image quality [2,43].

An approach focusing on obtaining cross-domain invariant representations was adopted, namely via learning light-invariant information or reflective representations in the zero-shot domain adaptation configuration, to address this challenge. Figure 1 displays the framework. A pair of well-lit images  $I^n$  and their respective artificially generated low-light images  $I^l$  are input into our framework. The pre-trained Retinex decomposition network (shown by the bottom gray block) is frozen during training and utilized primarily to infer pseudo-ground truths for reflectance and illumination,  $\hat{R}^l, \hat{R}^n, \hat{I}^l, \hat{I}^n$ , to oversee the reflectance decoder. Specifically, the pseudo-ground truths for illumination and the initial batch of reflectance forecasts,  $R_1^l, R_1^n$ , are passed into the proposed exchange–recomposition–consistency process in the right module to reconstruct and redecompose the second round of reflectance predictions,  $R_2^l, R_2^n$ , and to compute the redecomposition consistency loss  $\mathcal{L}_{rc}$ . The input images first undergo enhancement through a low-light



enhancement module (indicated by the purple dashed block) before being processed in the backbone network (light blue block) for global feature extraction  $g_f$  and local feature extraction  $g_b$ , producing richer and more semantically meaningful feature representations. These features are aligned using the Mutual Feature Alignment Loss  $\mathcal{L}_{mfa}$  (blue arrow). The CLIP and Dino models (represented as yellow and purple blocks, respectively) are utilized for extracting specific types of features. The reflectance features,  $R^l, R^n$ , are input into the reflectance decoder (light yellow block) to reconstruct the reflectance images  $\hat{R}^l, \hat{R}^n$  and compute the reflectance loss  $\mathcal{L}_{ref}$  and the redecomposition consistency loss  $\mathcal{L}_{rc}$  (indicated by brown and red labels, respectively). The first round of prediction results,  $R_1^l, R_1^n$ , are exchanged and the second round of predictions,  $R_2^l, R_2^n$ , is generated in the redecomposition process, where the redecomposition consistency loss  $\mathcal{L}_{rc}$  is calculated. Finally, the outputs for the detection tasks are processed through the neck and head module (green block). The implementation process is shown in Algorithm 1. The CLIP (ViT-based image encoder part) can learn powerful feature representations from large-scale multimodal data, exhibiting excellent generalization capabilities. Dino, through its unsupervised learning method of self-distillation, extracts robust image features from unlabeled data, enabling the model to maintain stable feature representations under varying lighting conditions.

Models can be trained effectively on well-lit images and be extended to low-light scenes through this approach, resulting in zero-shot day–night domain adaptation in object identification.

---

**Algorithm 1** Algorithm for framework outline of the methodology.

---

**Input:** Well-lit image  $I^n$ , artificially generated low-light image  $I^l$

**Output:** Refined reflectance predictions  $R_2^n, R_2^l$

1: **Initialization:**

Decompose  $I^n$  and  $I^l$  into reflectance and illumination:

$[L^n, R^n] \leftarrow \text{Retinex\_Decomposition}(I^n)$

$[L^l, R^l] \leftarrow \text{Retinex\_Decomposition}(I^l)$

2: **Step 1: Feature Extraction**

Extract features using backbone and auxiliary encoders:

$g_f, g_b \leftarrow \text{Backbone}(I^n, I^l)$  // Global and local features

$\text{clip\_features} \leftarrow \text{CLIP\_Image\_Encoder}(I^n, I^l)$  // Semantic features

$\text{dino\_features} \leftarrow \text{DINO}(I^n, I^l)$  // Self-supervised features

Combine features:

$\text{combined\_features} \leftarrow \text{Concatenate}(g_f, g_b, \text{clip\_features}, \text{dino\_features})$

Align all features using mutual feature alignment loss ( $\mathcal{L}_{mfa}$ ).

3: **Step 2: Reflectance Reconstruction**

Perform initial reflectance decoding:

$[R_1^n, R_1^l] \leftarrow \text{Reflectance\_Decoder}(\text{combined\_features})$

Apply exchange–recomposition process:

$[R_2^n, R_2^l] \leftarrow \text{Exchange\_Recomposition}(R_1^n, R_1^l)$

Compute redecomposition consistency loss ( $\mathcal{L}_{rc}$ ).

4: **Step 3: Final Output**

Return refined reflectance predictions:

5: **return**  $R_2^n, R_2^l$

---

### 3.1. Lighting Invariance Enhancement

Here, we revisit the Retinex theory [44], which posits that an image  $I$  can be decomposed into two constituents: reflectance  $R$  and illumination  $L$  ( $I = R \cdot L$ ). In this context, illumination affects the visibility of the image while reflectance remains constant. In zero-shot day–night domain adaptation for object detection, reflectance is considered an illumination-invariant counterpart, which is crucial for illumination-invariant detectors. Reflectance indicates that the learning module is fused to be used to enhance the detector's

capacity to adjust in low-light environments. The front end of the detector's backbone  $g_f$  is used to encode shallow information and branch out into a reflectance decoder. To stabilize the training of the decoder, the pre-trained Retinex network was utilized to generate pseudo-ground truths for reflectance and illumination. An illumination invariance enhancement scheme, which operates at the feature level, has been employed to enhance the detector's illumination invariance. To acquire feature representations that are not affected by changes in lighting, we stipulate that the output features  $F$  extracted from  $g_f$  and input into the reflectance decoder remains consistent between well-lit and low-light images. To achieve this, we employ a mutual feature alignment loss, which explicitly matches the features from well-lit ( $F^n$ ) and low-light ( $F^l$ ) conditions:

$$\mathcal{L}_{mfa} = \mathcal{KL}(F^n \parallel F^l) + \mathcal{KL}(F^l \parallel F^n) \quad (1)$$

where  $\mathcal{KL}(\cdot \parallel \cdot)$  represents the Kullback–Leibler (KL) divergence,  $F^n$  and  $F^l$  were the features from well-lit and low-light images extracted by  $g_f$ . These features were flattened and spatially averaged before being used to calculate the loss.

### 3.2. Low-Light Image Reconstruction

An image decomposition process is employed to further enhance reflectance learning. Given a set of low-light image  $I_l$  and a well-lit image  $I_n$ , a typical Retinex-based image decomposition algorithm [2,10] decomposes them into their respective reflectance and illumination, i.e., low-light reflectance  $R_1^l$  and illumination  $L^l$  for  $I_l$ , and well-lit reflectance  $R_1^n$  and illumination  $L^n$  for  $I_n$ . Ideally, the low-light reflectance  $R_1^l$  and the well-lit reflectance  $R_1^n$  should be interchangeable, and when combined with their respective illuminations,  $L^n$  and  $L^l$ , they should be able to reconstruct  $I_n$  and  $I_l$ , respectively. Utilizing this interchangeability, we added a constraint to strengthen image decomposition and reflectance representation learning.

Initially, swap the reflectance between well-lit and low-light images and reconstruct the images, such that  $I_2^l = R_1^n \cdot L^l$ ,  $I_2^n = R_1^l \cdot L^n$ . Then, the reconstructed images undergo a second round of decomposition. Our DAI-Net focuses on learning the illumination-invariant part of the image (i.e., the reflectance); hence, we use the same reflectance decoding branch in DAI-Net to decompose the reflection  $R_2^n$  and  $R_2^l$  from the reconstructed images  $I_2^n$  and  $I_2^l$ . To ensure consistency across the two decomposition processes, we introduce a redecomposition coherence loss.

$$\mathcal{L}_{rc} = \left\| R_1^n - R_2^l \right\|_1 + \left\| R_1^l - R_2^n \right\|_1 \quad (2)$$

The redecomposition coherence loss, compared to basic penalty losses, more effectively leverages the interchangeability of reflectance. This enhancement increases both the stability and accuracy of reflectance representation learning. Utilizing this approach allows us to significantly enhance the learning of reflectance in low-light conditions and substantially improve overall detection performance.

### 3.3. CLIP (ViT-Based Image Encoder) Framework

CLIP is a pre-trained model that pairs images with text through contrastive learning, which can learn features in high-dimensional spaces that exhibit strong generalization capabilities. The image encoder part of the integrated pre-trained CLIP model is used for feature extraction to enhance the performance of the strong detector under different lighting conditions.

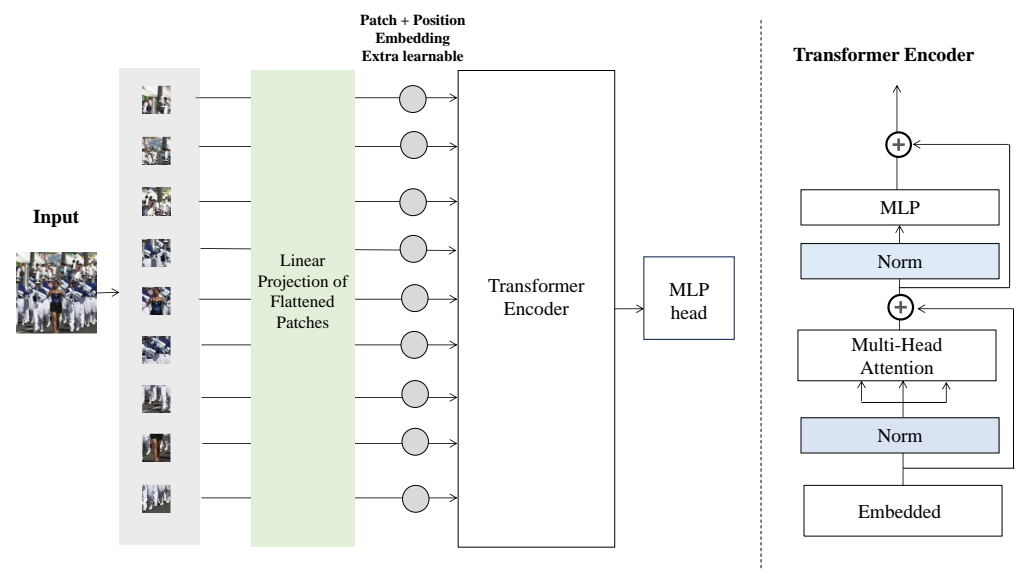
The image encoder of CLIP is a neural network designed to extract deep visual features from images. Typically, CLIP's image encoder is pre-trained on large-scale visual datasets,

enabling it to capture a wide range of visual concepts and object features. In principle, the CLIP (ViT-based image encoder part) employs either a convolutional neural network (CNN) or a vision transformer (ViT) as its primary architecture, depending on the specific version of the CLIP model. In this paper, we utilize the ViT as its main structure. As depicted in Figure 2, the ViT model processes images as sequences of patches embedded with positional information, which were fed into a transformer architecture for feature extraction. The self-attention mechanism within the ViT focuses on the significance of each image patch; its self-attention architecture generates outputs based on queries, keys, and values—these three components were derived through linear transformations of the inputs:

$$Attention(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_K}}\right)V \quad (3)$$

The dimensionality of the keys is denoted as  $d_K$ , and similarly, the ViT model incorporates a multi-head attention mechanism, projecting  $n$  times, where  $d_K = d_V = d_{\text{mod } el}/h$ , with each projection calculated in parallel, enhancing computational efficiency. These networks, through extensive learning from a vast array of image data, were capable of recognizing and understanding the contents within images, ranging from simple textures and shapes to complex scenes and object interactions.

To simplify the training process, a CLIP (ViT-based image encoder part) is integrated, which can improve the detection performance by exploiting advanced visual features, thus showing better adaptability and accuracy when dealing with extreme lighting or visually varying scenarios, improving the robustness of the model and the generalization of the features.



**Figure 2.** Structural overview of CLIP (ViT-based image encoder part): Images are put into the framework, where they are first divided into fixed-size patches. These image patches are then flattened and transformed into fixed-length feature vectors through linear projection. After incorporating positional embeddings and learnable embeddings, these feature vectors are processed by the transformer encoder, generating richer and more semantically meaningful feature representations. Finally, the feature vectors processed by the transformer encoder are passed through a multi-layer perceptron head (MLP head) to produce outputs for specific tasks, such as classification or detection.

### 3.4. DINO Framework

To further enhance the model's performance under extreme lighting conditions, Dino (Self-Distillation with No Labels) models were integrated. Dino is a self-supervised learning



technique that leverages label-free knowledge distillation to conduct self-supervised pre-training of visual features, thereby enhancing the model's robustness and adaptability.

The Dino model's fundamental concept, as seen in Figure 3, is training a student network to replicate the output of a teacher network via self-supervised learning. The teacher network is generated using the exponential moving average (EMA) of the student network. In practice, input images are progressively processed by different student networks, gradually learning high-quality visual features.

In the Dino model, the outputs generated by the teacher and student networks are represented as probability distributions normalized by a softmax function. The specific formula for the output probability distribution of the teacher network is as follows:

$$P_t(x) = \frac{\exp(g_{\theta_s}(x)/\mathcal{T}_t)}{\sum_{k=1}^K \exp(g_{\theta_s}(x)/\mathcal{T}_t)} \quad (4)$$

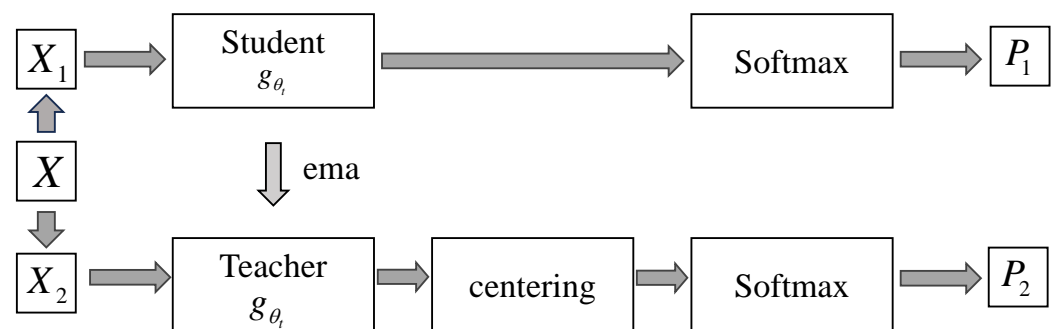
Student network output probability distribution:

$$P_s(x) = \frac{\exp(g_{\theta_s}(x)/\mathcal{T}_s)}{\sum_{k=1}^K \exp(g_{\theta_s}(x)/\mathcal{T}_s)} \quad (5)$$

where  $\mathcal{T}_s$  represents the temperature parameter of the student network's output, which controls the sharpness of the probabilistic distribution of the student network's output.  $\mathcal{T}_t$  denotes the temperature parameter of the teacher network's output, used to regulate the sharpness of the probabilistic distribution of the teacher network's output. The loss function is:

$$\min_{\theta_s} H(P_t(x), P_s(x)) = - \sum P_t(x) \log P_s(x) \quad (6)$$

Through this approach, the student network progressively aligns with the output of the instructor network to learn high-quality feature representations. Specifically, the Dino model learns high-quality features from unlabeled image data, which exhibit strong consistency under various lighting conditions, enhancing the model's generalization capabilities.



**Figure 3.** Structural overview of Dino: In this architecture, the input images  $X_1$  and  $X_2$  are individually input into the student network and the teacher network. Although both networks share the same architecture, their parameters are updated independently. Specifically, the student network processes the image  $X_1$ , generates features, and converts them into a probability distribution via a Softmax layer. The teacher network processes the image  $X_2$ , generates features, and converts them into a probability distribution  $P_2$  via a Softmax layer that includes a centering operation to eliminate bias in the predictions. The centering operation stops the gradient, ensuring that gradients are not back-propagated. During training, the instructor network's parameters are adjusted using exponential moving averages, ensuring that the teacher network provides a stable and accurate supervisory signal. The objective is to narrow the gap between the probability distributions generated by the student and teacher networks, hence improving the model's capacity to generalize.

### 3.5. Network Training

The network consists of two branches: detection and reflectance decoding. Regarding the former, the detection loss utilized by the selected detector is denoted as  $\mathcal{L}_{\text{det}}$ . The objective function for the reflectance decoding branch is comprised of three parts. The first segment has two planned losses:  $\mathcal{L}_{mfa}$  and  $\mathcal{L}_{rc}$ . The other two parts are the reflectance learning loss  $\mathcal{L}_{ref}$  and the image decomposition loss  $\mathcal{L}_{decom}$ . The loss function for the reflectance learning loss is defined as:

$$\mathcal{L}_{ref} = \text{MAE}(R, \hat{R}) + (1 - \text{SSIM}(R, \hat{R})) \quad (7)$$

The mean absolute error (MAE) measures the average absolute difference, while the structural similarity index measure (SSIM) evaluates the perceptual similarity between images. In practice,  $(R, \hat{R})$  is implemented as  $(R^l, \hat{R}^l)$  or  $(R^n, \hat{R}^n)$ . The image decomposition loss function is given by:

$$\mathcal{L}_{decom} = \mathcal{L}_{recon} + \lambda_{smooth} \mathcal{L}_{smooth} + \lambda_{ir} \mathcal{L}_{ir} \quad (8)$$

This loss function strengthens reflectance learning through image decomposition loss, which combines image reconstruction loss  $\mathcal{L}_{recon}$ , invariant reflectance loss  $\mathcal{L}_{ir}$ , and illumination smoothness loss  $\mathcal{L}_{smooth}$ . The  $\mathcal{L}_{recon}$  component aims to reconstruct the input image  $I$  from  $R \cdot \hat{L}$ , while  $\mathcal{L}_{smooth}$  and  $\mathcal{L}_{ir}$  are calculated between paired inputs. Specifically, the invariant reflectance loss  $\mathcal{L}_{ir}$  is defined as:

$$\mathcal{L}_{ir} = \text{MSE}(R^l, R^n) + (1 - \text{SSIM}(R^l, R^n)) \quad (9)$$

where MSE is the mean squared error, adding to the robustness of the loss function by assessing both error and similarity.

The overall loss function comprises these components to effectively enhance model performance through accurate reflectance reconstruction and regularization of illumination and reflectance properties across different lighting conditions.

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{mfa} \mathcal{L}_{mfa} + \lambda_{rc} \mathcal{L}_{rc} + \mathcal{L}_{ref} + \mathcal{L}_{decom} \quad (10)$$

Each term and its corresponding regularization constant were determined based on prior research and experimental validation:

The values for  $\lambda_{mfa} = 0.1$  and  $\lambda_{rc} = 0.001$  were determined through grid search experiments on the WIDER FACE validation set. The grid search aimed to identify values that optimized detection accuracy (mAP) while ensuring stable convergence of the training process.  $\lambda_{smooth}$  and  $\lambda_{ir}$  are constants that were inherited from the Retinex-based image decomposition model in [10], where  $\lambda_{smooth} = 0.5$  and  $\lambda_{ir} = 0.01$  have been widely used to maintain reflectance consistency and illumination smoothness.

## 4. Experiment

In this section, the problem of face detection in dark environments is mainly discussed, and the performance of the proposed method is evaluated in detail under extreme lighting conditions.

### 4.1. Datasets and Evaluation Indicators

#### 4.1.1. Datasets

WIDER FACE is chosen as the source domain with good illumination, and the trained model is tested on the target domain DARK FACE. These domains are denoted as WIDER

FACE → DARK FACE (from source to target). The WIDER FACE dataset, which was first released in 2015 (version 1.0), contains a total of 32,203 images selected from the WIDER dataset, with annotations for 393,703 faces. Additionally, each face annotation is accompanied by more detailed information covering various scenes, such as daily life, outdoor activities, parties, and streets, with a range of lighting conditions, including both good and insufficient lighting, as shown in Figure 4. In contrast, the DARK FACE dataset focuses on face detection tasks in low-light environments and includes 6000 annotated images as well as 9000 non-annotated images. This dataset aims to improve algorithm performance for face detection under poor lighting conditions, as shown in Figure 5.



Figure 4. WIDER FACE.



Figure 5. DARK FACE.

#### 4.1.2. Evaluation Indicators

In this study, the mean average precision (mAP) was used as the primary metric to evaluate the performance of the model. mAP is a commonly used performance measure in object detection tasks, as it comprehensively takes into account the model's precision and recall, thereby providing a thorough evaluation of the model's detection capability. Precision measures the proportion of samples predicted to be positive that are actually positive. The formula is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

where  $TP$  represents the number of true examples, and  $FP$  represents the number of false positive examples. Recall measures the proportion of samples that are actually positive and are correctly predicted to be positive. The formula is:

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

where  $FN$  denotes the number of false negative cases. Average precision (AP) is the area under the precision–recall curve, calculated at different thresholds. It evaluates a model's performance by summarizing the trade-off between precision and recall across these thresholds. mAP is for all classes of The average value of average precision, calculated as:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (13)$$

where  $N$  is the total number of categories and  $AP_i$  is the average accuracy of the  $i$ -th category.

#### 4.2. Ablation Experiment

In this section, the various modules of the model are analyzed in depth through ablation experiments, with special attention paid to the synergy between the CLIP (ViT-based image encoder part) and the Dino self-supervised learning module. At the same time, we have changed the loss function to improve the performance of the model, but the effect is not ideal. We evaluate the impact of these modules on the overall performance of the model by removing them one by one. The experimental results, as shown in Table 1, indicate that when CLIP (ViT-based image encoder part) or Dino are used alone, the performance of the model decreases due to the limitations of both. The average accuracy mean values were 26.8% and 25.37%, respectively. However, when the two are combined, the overall performance of the model is significantly improved, with an average accuracy value of 29.6%. The reason is as follows:

**Table 1.** Ablation experiment.

Method	mAP (%)
Dainet + dino	25.37
Dainet + CLIP (ViT-based image encoder part)	26.81
Dainet + CLIP (ViT-based image encoder part) + dino	29.6

First of all, the CLIP (ViT-based image encoder part) is based on vision transformer (ViT), which can effectively capture the global semantic information of the image. It has unique advantages in overall scene understanding, especially when switching between day and night scenes. By capturing global features, it can improve the global semantic expression ability of the model. However, the ability of CLIP (ViT-based image encoder part) to capture local details is relatively limited, and there are some shortcomings in fine-grained feature extraction. This limitation stems from the pre-training objective of the CLIP image encoder, which prioritizes extracting high-level global information while lacking optimization for pixel-level or region-specific details. In low-light object detection tasks, this limitation becomes particularly pronounced, as the model needs to accurately capture subtle variations in the image to adapt to complex lighting conditions.

Secondly, as a self-supervised learning framework, the Dino module is good at capturing local details in images. Especially in scenes with significant changes in lighting conditions, Dino can learn detailed features related to lighting changes well. Although Dino has advantages in local feature extraction, when used alone, it faces limitations due to the lack of explicit supervision and global semantic alignment mechanisms. Its global feature modeling ability is relatively weak, making it challenging to comprehensively represent the overall contextual information of images.

When CLIP (ViT-based image encoder part) and Dino module are introduced at the same time, their complementarity in feature extraction is fully exerted. CLIP (ViT-based image encoder part) provides powerful global semantic features, ensuring that the model can obtain comprehensive structural information in complex scenarios; Dino strengthens the capture of local details, especially in scenes with drastic lighting changes. Dino enhances the model's perception of key details through fine-grained feature learning. The combination of the two not only improves the overall feature representation ability of the model, but also enhances the robustness of the model in day and night scene switching.

### 4.3. Comparative Experiment

In the case of zero-shot domain adaptation, the dark scene is taken as the target domain. The zero-shot day–night domain adaptation method [2,45] is directly applied to DSFD detection, and this method is compared with CIConv [45], Sim-MinMax [43], and DAI-Net [13].

We selected CIConv, Sim-MinMax, and DAI-Net as the objects of comparison in our experiments for the following reasons: Firstly, CIConv is a traditional low-light detection method based on local convolution, and its performance reflects the limitations of classical approaches in day–night scenarios. Secondly, Sim-MinMax is a domain adaptation method that achieves domain transfer by minimizing image differences between the source and target domains, serving as a representative benchmark for cross-domain detection. Lastly, DAI-Net is a network specifically designed for low-light scenarios, equipped with multi-scale feature extraction and illumination variation handling capabilities, representing the state-of-the-art in low-light object detection tasks.

The main limitation of CIConv is its dependence on local convolution, difficulty in capturing significant global lighting differences between day and night images, and insufficient ability to handle details under complex lighting conditions. Although Sim-MinMax achieves domain adaptation by minimizing the image difference between the source domain and the target domain, it is not robust enough to drastic lighting changes during day and night, resulting in insufficient generalization performance when dealing with cross-domain scenes, especially in low-light environments. The performance is weak. In addition, Sim-MinMax lacks effective capture of fine-grained features, making it difficult for the model to accurately identify details in scenes with drastic lighting changes. The DAI-Net network demonstrates excellent domain migration ability through effective multi-scale feature extraction and low-light processing, especially when dealing with changes in lighting conditions during the day and night, it can better capture and adapt to face features. This lays the foundation for the robustness of the model in cross-domain tasks. In order to further improve the performance of DAI-Net in the day–night adaptive face detection task in the zero-sample domain, we introduce the CLIP (ViT-based image encoder part) and the Dino self-supervised learning module based on DAI-Net. By adding CLIP and Dino, the model has stronger global and local feature extraction capabilities. CLIP (ViT-based image encoder part)s are good at capturing the global semantic information of images, while Dino effectively captures local details and domain invariance features through self-supervised learning. Through self-supervised learning, Dino can learn cross-domain feature representations to deal with domain migration problems under different lighting conditions such as day and night. Combined with the global features provided by the CLIP (ViT-based image encoder part), Dino can effectively integrate the learned domain invariant features with global context information, further enhancing the domain-adaptive ability of the model. Based on multi-scale feature representation theory, the diversification of feature space helps to improve the generalization ability of the model. In our approach, the global features of CLIP are combined with the local features of Dino to enrich the feature representation space of the model. This multi-scale feature fusion strategy enables the model to better capture important information related to face detection from different scales, especially in the complex environment of day and night scene, the model shows higher robustness and generalization ability. Moreover, the Retinex-based reflectance representation learning module adds another layer of adaptability by disentangling reflectance from illumination through supervised and semi-supervised decomposition processes. This module ensures that the model learns illumination-invariant representations, which remain consistent under various lighting conditions. These representations allow the model to focus on the stable, intrinsic features of the scene, enabling it to handle diverse lighting

scenarios effectively. Based on multi-scale feature representation theory, the diversification of feature space helps to improve the generalization ability of the model. In our approach, the global features of CLIP are combined with the local features of Dino to enrich the feature representation space of the model. This multi-scale feature fusion strategy enables the model to better capture important information related to face detection from different scales, especially in the complex environment of day and night scenes. The proposed framework demonstrates its adaptability in diverse scenarios by capturing both global and local features robustly, excelling particularly in low-light domains where precise feature isolation is critical. The model shows higher robustness and generalization ability in such challenging environments.

We compare the new model with the basic DAI-Net. Experimental results, as shown in Table 2, show that after adding CLIP (ViT-based image encoder part) and Dino, the performance of the model in diurnal adaptive tasks is significantly improved.

**Table 2.** WIDER FACE → DARK FACE test set using DSFD.

Method	mAP (%)
CICov [45]	18.4
Sim-MinMax [43]	25.7
DAI-Net [13]	28.0
ours	29.6

On the whole, these methods have certain limitations in the zero-sample day and night adaptive task, which are mainly reflected in the insufficient ability to deal with illumination changes, the imbalance between global and local feature extraction, and the lack of domain-adaptive ability. Therefore, it is difficult for them to achieve good generalization results in complex cross-domain tasks. In contrast, our model achieved an average accuracy of 29.6% in dark domain scenes, 11.2% higher than CICov, 3.9% higher than Sim-MinMax, and 1.6% higher than DAI-Net, as shown in Table 2, demonstrating stronger robustness and domain adaptability, achieving optimal performance in current tasks.

Based on the findings from the ablation and comparative experiments, we can draw the following conclusions: The ablation experiments reveal the individual limitations of the CLIP image encoder and Dino when used in isolation, underscoring the necessity of complementary techniques. The comparative experiments further highlight the performance enhancements achieved by integrating these two components within the DAI-Net framework. These consistent improvements in low-light object detection tasks validate the critical role of leveraging both global and local features. Together, these findings provide compelling evidence of the framework's effectiveness and its superiority in addressing zero-shot day–night domain adaptation challenges.

#### 4.4. Visualization of Results

Figure 6 illustrates the results of face detection in dark environments. The upper row displays images that have been manually brightened, while the lower row presents the outcomes of face detection in the dark using the model architecture described in this study. In the lower row, recognized faces are highlighted in yellow. Through comparative analysis, it is evident that the dark domain detection method in this paper can still provide stable and reliable detection results under low-light conditions, further demonstrating its effectiveness and robustness in extreme lighting conditions.





Figure 6. Visualization of results.

## 5. Conclusions

This paper addresses object detection in dark environments under the novel zero-shot domain adaptation (ZSDA) setting for dark domains. We extended DAI-Net by integrating the image encoder from the CLIP and Dino models, significantly enhancing performance in extreme lighting conditions.

The integration of the CLIP model (ViT-based image encoder) and Dino model introduces advanced visual features and highly consistent representations, ensuring robust feature extraction across various lighting environments. Experimental results on the DARK FACE dataset demonstrate that these enhancements greatly improve the model's generalization and resilience, enabling accurate object detection in complex lighting scenarios. These findings validate the effectiveness and superiority of the proposed approach.

Future research could explore applying the framework to more complex domain shift scenarios, such as adverse weather conditions, extreme occlusion, or domain variations caused by sensor differences, to validate its adaptability and robustness. Additionally, while this study focuses on object detection in dark environments, the framework shows potential for expansion to other tasks, including semantic segmentation, instance segmentation, and video object tracking under complex lighting conditions, further broadening its applicability.

**Author Contributions:** Conceptualization, H.S., and Y.L.; methodology, H.S., and Y.L.; software, P.Z.; validation, Y.L.; formal analysis, H.S., and Y.L.; investigation, Y.L.; resources, H.S.; data curation, Y.L.; writing—original draft preparation, H.S., and Y.L.; writing—review and editing, Y.L., and Z.C.; visualization, P.Z.; supervision, H.S.; project administration, P.Z.; funding acquisition, H.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Harbin City Science and Technology Plan Projects grant number 2022ZCZJCG006, the Science and Technology Collaborative Innovation Project in Heilongjiang Province grant number LJGXCG2023-097.

**Data Availability Statement:** The data supporting the results reported in this article are available from publicly accessible sources: WIDER Face dataset for training and validation images, accessible at <http://shuoyang1213.me/WIDERFACE/> (accessed on 24 July 2024), including annotations for the training and validation sets. Dark Face dataset for testing samples, accessible at <http://www.ug2challenge.org/> (accessed on 24 July 2024), which provides benchmarks and resources for low-light face detection.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1254.
2. Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; Zhang, Y. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 12504–12513.
3. Guo, C.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.; Cong, R. Zero-reference deep curve estimation for low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1780–1789.
4. Wang, W.; Xu, Z.; Huang, H.; Liu, J. Self-aligned concave curve: Illumination enhancement for unsupervised adaptation. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa Portugal, 10–14 October 2022; pp. 2617–2626.
5. Wang, W.; Yang, W.; Liu, J. Hla-face: Joint high-low adaptation for low light face detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16195–16204.
6. Wu, W.; Weng, J.; Zhang, P.; Wang, X.; Yang, W.; Jiang, J. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5901–5910.
7. Cui, Z.; Qi, G.-J.; Gu, L.; You, S.; Zhang, Z.; Harada, T. Multitask AET with orthogonal tangent regularity for dark object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2553–2562.
8. Loh, Y.P.; Chan, C.S. Getting to know low-light images with the exclusively dark dataset. *Comput. Vis. Image Underst.* **2019**, *178*, 30–42. [[CrossRef](#)]
9. Yang, S.; Luo, P.; Loy, C.-C.; Tang, X. Wider face: A face detection benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5525–5533.
10. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep retinex decomposition for low-light enhancement. *arXiv* **2018**, arXiv:1808.04560.
11. Guo, J.; Deng, J.; Lattas, A.; Zafeiriou, S. Sample and computation redistribution for efficient face detection. *arXiv* **2021**, arXiv:2105.04714.
12. Li, J.; Wang, Y.; Wang, C.; Tai, Y.; Qian, J.; Yang, J.; Wang, C.; Li, J.; Huang, F. DSFD: Dual shot face detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5060–5069.
13. Du, Z.; Shi, M.; Deng, J. Boosting Object Detection with Zero-Shot Day-Night Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 12666–12676.
14. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR: Cambridge, MA, USA, 2021; pp. 8748–8763.
15. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9650–9660.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part I; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
17. Redmon, J. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
18. Tian, Z.; Chu, X.; Wang, X.; Wei, X.; Shen, C. Fully convolutional one-stage 3D object detection on lidar range images. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 34899–34911.
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
20. Dai, D.; Van Gool, L. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 3819–3824.
21. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-shot multi-level face localisation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5203–5212.
22. Hu, P.; Ramanan, D. Finding tiny faces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 951–959.
23. Liu, Y.; Shi, M.; Zhao, Q.; Wang, X. Point in, box out: Beyond counting persons in crowds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6469–6478.

24. Liu, Y.; Tang, X. BFBBox: Searching face-appropriate backbone and feature pyramid network for face detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13568–13577.
25. Ming, X.; Wei, F.; Zhang, T.; Chen, D.; Wen, F. Group sampling for scale invariant face detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3446–3456.
26. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
27. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
28. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
29. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Mallocci, M.; Kolesnikov, A.; et al. The Open Images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vis.* **2020**, *128*, 1956–1981. [[CrossRef](#)]
30. Liu, R.; Ma, L.; Zhang, J.; Fan, X.; Luo, Z. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10561–10570.
31. Sasagawa, Y.; Nagahara, H. YOLO in the dark-domain adaptation method for merging multiple models. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part XXI; Springer: Berlin/Heidelberg, Germany, 2020; pp. 345–359.
32. Mo, Y.; Han, G.; Zhang, H.; Xu, X.; Qu, W. Highlight-assisted nighttime vehicle detection using a multi-level fusion network and label hierarchy. *Neurocomputing* **2019**, *355*, 13–23. [[CrossRef](#)]
33. Vankadari, M.; Garg, S.; Majumder, A.; Kumar, S.; Behera, A. Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part XXVIII; Springer: Berlin/Heidelberg, Germany, 2020; pp. 443–459.
34. Liu, Y.; Wang, F.; Deng, J.; Zhou, Z.; Sun, B.; Li, H. MoGFace: Towards a deeper appreciation on face detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4093–4102.
35. Hashmi, K.A.; Kallempudi, G.; Stricker, D.; Afzal, M.Z. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 6725–6735.
36. Liu, M.-Y.; Tuzel, O. Coupled generative adversarial networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
37. Wang, Q.; Breckon, T.P. Generalized zero-shot domain adaptation via coupled conditional variational autoencoders. *Neural Netw.* **2023**, *163*, 40–52. [[CrossRef](#)] [[PubMed](#)]
38. Gao, H.; Guo, J.; Wang, G.; Zhang, Q. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9913–9923.
39. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
40. Deng, X.; Wang, P.; Lian, X.; Newsam, S. NightLab: A dual-level architecture with hardness detection for segmentation at night. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16938–16948.
41. Sakaridis, C.; Dai, D.; Gool, L.V. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 7374–7383.
42. Du, Z.; Deng, J.; Shi, M. Domain-general crowd counting in unseen scenarios. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 561–570. [[CrossRef](#)]
43. Luo, R.; Wang, W.; Yang, W.; Liu, J. Similarity min-max: Zero-shot day-night domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 8104–8114.
44. Land, E.H. The retinex theory of color vision. *Sci. Am.* **1977**, *237*, 108–129. [[CrossRef](#)] [[PubMed](#)]
45. Lengyel, A.; Garg, S.; Milford, M.; van Gemert, J.C. Zero-shot day-night domain adaptation with a physics prior. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4399–4409.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.