# Watching it in Dark: A Target-aware Representation Learning Framework for High-Level Vision Tasks in Low Illumination

Yunan Li[1,2,3], Yihao Zhang[1,2], Shoude Li[1,2], Long Tian[1], Dou Quan[4], Chaoneng Li[5], and Qiguang Miao[⋆1,2,3]

[1] School of Computer Science and Technology, Xidian University, Xi'an, China
[2] Xi'an Key Laboratory of Big Data and Intelligent Vision, Xi'an, China
[3] Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, Xi'an, China
[4] School of Artificial Intelligence, Xidian University, Xi'an, China
[5] School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou, China

`{yunanli, tianlong, quandou, qgmiao}@xidian.edu.cn`, `{zhangyihao,shoulderlee}@stu.xidian.edu.cn`, `chaonengli@lzjtu.edu.cn`

**Abstract.** Low illumination significantly impacts the performance of learning-based models trained under well-lit conditions. While current methods mitigate this issue through either image-level enhancement or feature-level adaptation, they often focus solely on the image itself, ignoring how the task-relevant target varies along with different illumination. In this paper, we propose a target-aware representation learning framework designed to improve high-level task performance in low-illumination environments. We achieve a bi-directional domain alignment from both image appearance and semantic features to bridge data across different illumination conditions. To concentrate more effectively on the target, we design a target highlighting strategy, incorporated with the saliency mechanism and Temporal Gaussian Mixture Model to emphasize the location and movement of task-relevant targets. We also design a mask token-based representation learning scheme to learn a more robust target-aware feature. Our framework ensures compact and effective feature representation for high-level vision tasks in low-lit settings. Extensive experiments conducted on CODaN, ExDark, and ARID datasets validate the effectiveness of our approach for a variety of image and video-based tasks, including classification, detection, and action recognition. Our code is available at `https://github.com/ZhangYh994/WiiD`.

**Keywords:** Representation learning · Day-night adaptation · Target-aware masking · Contrastive learning

## 1 Introduction

To boost the performance in low light, a naive solution is to enhance the illumination before passing the images to the downstream networks [3,11,30,32,35,41].

⋆ Corresponding author

However, these low-level enhancements often rely on pixel-level objectives like L1 or MSE losses. These losses always require paired data and may introduce noise, as pixel-wise mapping disregards semantic information in the scene [46, 54, 56]. Therefore, some researchers attempt to solve this problem from the view of whole image and introduce transfer learning techniques [2, 8, 37, 39, 53] and domain adaptation [22, 49, 50] to push the distribution of low-lit images to well-lit ones from image-level and feature-level, respectively. Although these methods yield appealing results, it is still worth pointing out that the substantial domain gap in this area presents a significant challenge for adaptation. The unidirectional transferring is associated with a rapid change in the amount of available information. When adapting from well-lit to low-lit conditions, maintaining consistency in data distribution often results in the loss of significant color and texture information, which is vital for high-level vision tasks.

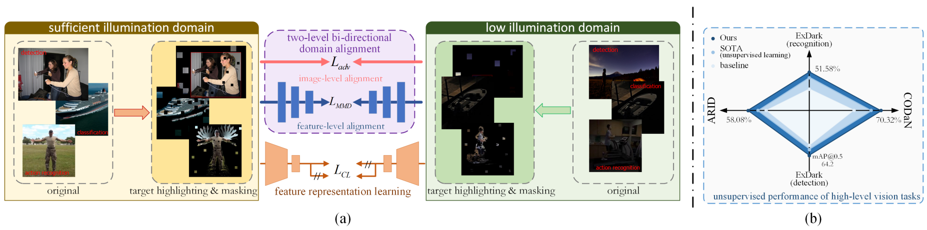



**Fig. 1:** Principle diagram and performance matrix of our framework. (a) The principle of our framework. (b) The performance comparisons on different high-level vision tasks.

The process of dark-to-bright adaptation can be considered as an entropy production process, which involves predicting high-information content from low-information data. This can be modeled with a linear approach as described in [24]. The problem is ill-posed because the solution is not unique [17]. Unlike other image degradation problems such as dehazing [27, 28] and deraining [55], dark-to-bright adaptation lacks a physical model to constrain the transformation, leading to unexpected noise in the optimization process. A more effective approach involves defining an intermediate state and aligning the dark and bright domains toward this state from both directions. This alignment should consider both low-level and high-level information, as high-level tasks primarily focus on semantic information rather than purely visual details.

Meanwhile, in contrast to low-level vision primarily focuses on enhancing visual quality such as more vivid colors or clearer textures, high-level vision is more concerned about the semantic understanding of the scene, which is related to the target entity for the task. For example, tasks like object recognition/detection/tracking are concerned with identifying the class or location of the target object [22, 33], while action recognition focuses on analyzing the motion of performers in a video sequence [26, 29, 54]. Previous research [47] has demonstrated that the region relevant to the target is consistently emphasized and corresponds to a higher response in the feature map when trained with a network. Moreover, the presence of artificial light sources such as streetlights, introduces non-uniform illumination in dark scenes. Consequently, the degradation for the target regions may vary even in one image depending on their distance to the light source. If adaptation is conducted based on the entire image, the features of the target region may be not learned effectively, potentially hindering the ultimate semantic vision tasks.

Based on these concerns, we propose a target-aware representation learning framework for high-level vision tasks. As depicted in Fig.1, considering the challenge caused by the huge gap between the dark and bright domains, we design a two-level bi-directional alignment strategy based on unpaired data. The low illumination images are first brightened as per some enhancement priors to reveal visual details whereas the well-lit ones are darkened via network training. Then these two domains are aligned from both image-level and feature-level via adversarial learning and Maximum Mean Discrepancy (MMD) constraint toward an intermediate state, facilitating the subsequent representation learning. To ensure the feature representation refinement is focused on the target rather than the entire scene, we design a target-aware self-supervised learning (SSL) paradigm. It first leverages the saliency theory for static attention and the Temporal Gaussian Mixture Model (TGMM) to highlight the target's movement and then uses different masking strategies for target and non-target regions to balance the diversity of samples and the integrity of the target. Finally, contrastive learning is employed to enable generalizable high-level models. This framework can be a common solution for high-level tasks in low-illumination environments and can be embedded in any high-level tasks.

The contributions of our method can be summarized as three-fold:

1. We propose a two-level bi-directional domain alignment scheme to address the huge domain gap challenge, which cannot be adequately mitigated through one-way adaptation. This scheme provides an intermediate state that alleviates noise generation while retaining prediction-relevant information. The alignment process is achieved at both the image and feature levels, leveraging adversarial learning and MMD constraints, respectively.
2. We design a target-aware masked contrastive learning strategy enhancing feature representation for high-level vision tasks. we employ the saliency theory and TGMM to highlight the location and motion of the target, and then leverage distinct masking strategies for target and non-target regions. Subsequently, contrastive learning is employed to promote a more robust feature representation.
3. As a universal representation learning framework, our model can be integrated into various high-level vision tasks. Experiments on three public datasets validate the effectiveness of our design in the tasks of object recognition, detection, and action recognition in low illumination, and demonstrate the superior performance compared to state-of-the-art methods.

## 2 Related Works

### 2.1 Low Illumination Enhancement Meets High-level Vision Tasks.

Researchers have extensively discussed strategies for generalizing daytime data-trained high-level vision models to dark scenes, such as image classification [22], object detection [33], and action recognition [54]. Traditional solutions adopt a two-stage approach, integrating the enhancement procedure with the downstream vision models. Early enhancement techniques, encompassing histogram equalization [43] and Retinex theory [21], are simple handicraft feature-based image processing approaches. Then the emergence of day-night datasets [6, 12, 48] containing pairs of low-lit and well-lit images has enabled deep learning-based enhancement methods [3, 11, 30, 32, 35]. However, the task-specific paired data is not always available. Moreover, these pixel-to-pixel enhancements tend to overfit to the specific data distribution of the training set and easily introduce some noise, limiting their applications in other scenes. Then unsupervised image translation techniques [36, 39, 53] offer an alternative approach for day-night image transfer.

### 2.2 Day-night Domain Adaptation

Since improved visibility does not always mean better performance for high-level tasks [31], many studies [1, 8, 60] have turned to domain adaptation, aiming to minimize the distribution gap between the two illumination domains from the feature level. Sasagawa *et al.* [40] design a glue layer for learning the latent features of two domains. Lengyel *et al.* [22] introduce domain adaptation using physical induction priors to capture illumination-robust features. Wang *et al.* [50] introduce a high-low level domain adaptation method for face detection in dark, where the low-level adaptation focuses on pixel-wise brightening and the high-level adaptation enhances synthesized features. Luo *et al.* [34] present a zero-shot day-night domain adaptation approach to free the requirement of task-specific low-light data by synthesizing it from well-lit data, and then they perform a min-max similarity optimization from the image and feature level, respectively. In contrast to unidirectional domain adaptation methods, our approach performs a two-level bidirectional domain alignment. The compromise transferring strategy retains the prediction-related information and mitigates noise, facilitating the downstream high-level vision tasks. Unlike the recent two-level methods [34, 50], the intermediate state is not only built by the low-level adaptation but yielded by ensuring consistency at both the image and feature levels, reducing the potential error accumulation. Moreover, the alignment on both levels is achieved by maximizing domain-wise similarity, mitigating potential divergence issues associated with no effective constraints owing to the minimization of similarity.

### 2.3 Masked Contrastive Learning

Masked Autoencoder (MAE) [9, 13, 42] and contrastive learning [5, 10, 14] are two unsupervised representation learning techniques. MAE trains the encoder by masking parts of the input data and then reconstructing them, while contrastive learning refines feature representation by comparing similar (positive) and dissimilar (negative) pairs of data points. Some studies combine the advantages of mask-based techniques with contrastive learning to enhance representation learning. Yao *et al.* [57] establish multiple branches and leverage contrastive learning between both the masked and unmasked sub-branches. [18, 62] employ two identical networks to extract features from both masked and original images, respectively, and conducts contrastive learning between them after reconstructing the masked portions. By contrast, we do not merely transplant the masking strategy into contrastive learning. Given our focus on high-level vision tasks, we begin by highlighting the task-relevant targets. Then distinct masking strategies are employed, determined by whether the region pertains to the target or non-target. This scheme helps to filter out interference from task-irrelevant factors during the representation learning. Compared to the random masking in MAE and the image-based view augmentation in traditional contrastive learning, our method ensures learning of a robust feature representation of target itself rather than other regions in the image such as the background, and it facilitates downstream tasks more effectively.

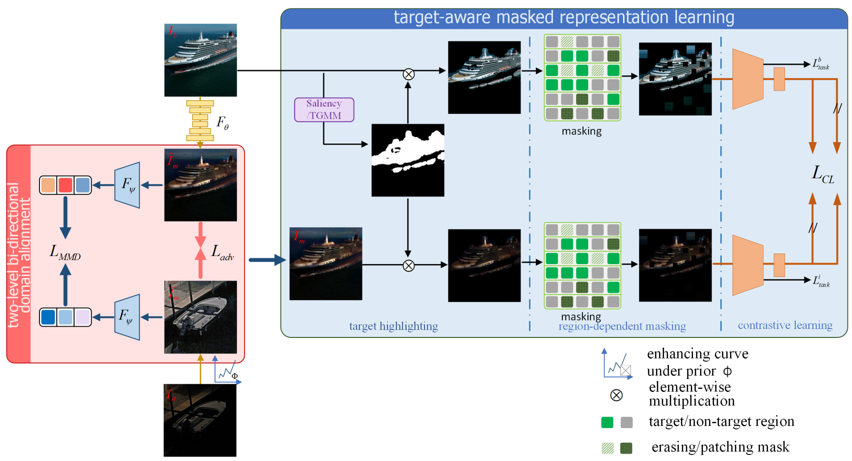



**Fig. 2:** Overview of our method. We first obtain approximations of the intermediate state $I_m$ from the low-lit input $I_d$ and its well-lit counterpart $I_b$ by enhancing and darkening, respectively. Then they are aligned using $\mathcal{L}_{adv}$ and $\mathcal{L}_{MMD}$ at the image and feature levels simultaneously. Following the bi-directional alignment, we employ a target-aware masking representation learning scheme, which begins with saliency/TGMM-driven target region highlighting. To increase the diversity of target feature representation, distinct masking strategies are employed to erase/reveal the target/non-target regions, respectively. Finally, a non-negative contrastive learning scheme is utilized to further refine the feature representation.

# 3 Proposed Method

## 3.1 Overview

The pipeline of our target-aware representation learning framework is depicted in Fig.2. As our primary concern lies in addressing the significant gap cannot solved by unidirectional day-night domain adaptation, we first present the bidirectional alignment of low-lit and well-lit domains. Using prior-driven image adjustment parameters, the input low-lit image $I_d$ is first enhanced to derive its approximation of intermediate state $\tilde{I}_m$. Simultaneously, the well-lit input is fed into a network to learn its prediction of the intermediate state $\hat{I}_m$. Then these two domains are aligned via adversarial loss $\mathcal{L}_{adv}$ and MMD loss $\mathcal{L}_{MMD}$ to achieve consistency at both low and high levels. After that, the semantic task-relevant target region is highlighted based on saliency theory and the Temporal Gaussian Mixture Model, depending on whether the input is an image or a video, respectively. Different masking strategy is then utilized to provide a more diverse representation of the target. Finally, a better feature representation of target is learned through the contrastive learning scheme to benefit the low illumination semantic vision task.

### 3.2 Two-level Bi-directional Domain Alignment

The significant gap between insufficient and sufficient illumination domains and the differences between pixel-level and semantic-level features hamper the performance on high-level vision tasks in the context of unidirectional day-to-night domain adaptation. Therefore, it is necessary to learn a more sophisticated way of transfer learning across two illumination domains to make them meet together at both the image level and feature level.

**Image-level alignment.** Assuming we have a low-lit input $I_d$ and a well-lit one $I_b$, the image-level alignment aims to find an intermediate state $I_m$ to ensure their consistency in image appearance. It represents a compromise strategy, which neither results in information loss due to over-darkening nor introduces noise by attempting to recover information that has already been lost.

As $I_m$ is not determined, there is no domain constraint for aligning. Then we can find inspiration from general visual enhancement for the mapping of $I_d \rightarrow I_m$. Such an enhancement can be deemed a prior-driven parameter conditioning. For instance, a well-lit image is usually considered with adequate brightness and vivid color. The enhancement is to adjust these parameters for a better view. Therefore, We argue that there is a set of low-lit image enhancing parameters $\Phi = \{\phi_1, \phi_2, \cdots, \phi_n\}$, then the enhancement can be expressed as $\tilde{I}_m = E(I_d|\Phi)$, where $\tilde{I}_m$ is the approximation of $I_m$, and $E(\cdot)$ refers to the enhancing function, namely the adjusting curve derived under the priors $\Phi$.

As the darkening process is not common, it is hard to find prior as guidance. Therefore, we train a network $\hat{I}_m = F_\theta(I_b)$ to achieve it, where $F_\theta$ is an encoder-decoder network parameterized by $\theta$. Then image-level alignment should be conducted between $\tilde{I}_m$ and $\hat{I}_m$. However, as we employ unpaired $I_d$ and $I_b$, the pixel-to-pixel constraints like [50] do not fit our requirement. Inspired by [36], we use adversarial learning under cycle-consistency to achieve it:

$$\mathcal{L}_{adv} = \mathbb{E}_{I_d \sim P_{data(I_d)}}[\log D(E(I_d|\phi))] + \mathbb{E}_{I_b \sim P_{data(I_b)}}[\log(1 - D(F_\theta(I_b)))]. \quad (1)$$

**Feature-level alignment.** Although the image-level alignment alleviates the gap problem of unidirectional adaptation, the semantic consistency of features should also be considered since we aim at dealing with high-level vision tasks. Similarly, as there is no constraint that the two domains are for the same task, $\tilde{I}_m$ and $\hat{I}_m$ not only differ from each other in content but also lead to various high-level features. Thus we need to consider drawing them together from the view of illumination in a higher dimension. Then we turn to the Maximum Mean Discrepancy (MMD) as the optimization objective, being expressed as:

$$L^2_{\mathrm{MMD}}(\tilde{I}_m, \hat{I}_m) = \left\| \frac{1}{N_d} \sum_{i=1}^{N_d} \phi(F^i_\psi(\tilde{I}_m)) - \frac{1}{N_b} \sum_{j=1}^{N_b} \phi(F^j_\psi(\hat{I}_m)) \right\|^2_{\mathcal{H}}, \tag{2}$$

where $F_\psi$ is a pre-trained network for extracting high-dimension features. $L^2_{\mathrm{MMD}}$ represents the squared MMD loss in the reproducing kernel Hilbert space (RKHS) endowed with the mapping function of $\phi(\cdot)$, which is usually a set of radial basis functions. On the right side of the equation, the first term calculates the mean of the mapped features from $\tilde{I}_m$ in the RKHS upon $N_d$ samples, and the second one is for the mean of features from the $\hat{I}_m$ in the RKHS upon $N_b$ samples.

By minimizing $\mathcal{L}_{MMD}$, the mean discrepancy between two approximations of the intermediate state $I_m$ in the new feature space is reduced and a real consistent $I_m$ is derived, thereby promoting the alignment of features. This step is crucial for reducing the distributional differences from the feature level, helping the high-level feature become illumination-independent.

### 3.3 Target-aware Masked Representation Learning

Although the two-level bidirectional alignment can bridge domains with different illumination conditions, it is still a kind of globally coarse-grained alignment. For any high-level tasks, the target entity is always the essence. As shown in Fig.3(a) and (b), the illumination can vary significantly across different locations within a single scene, leading to changes in the appearance of a target. Therefore, attention should be paid to the target itself rather than the entire image for feature representation learning schemes.

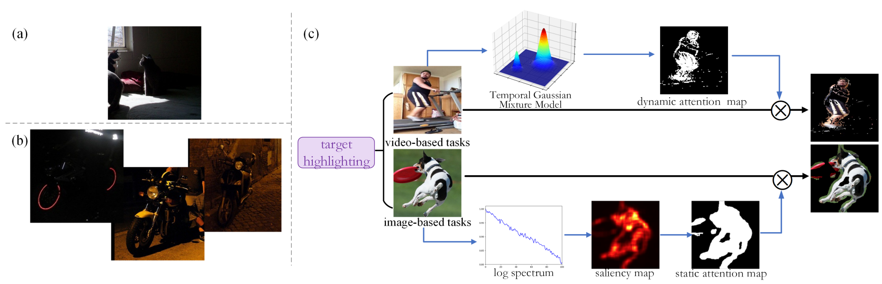



**Fig. 3:** The reason and structure of target highlighting. (a) Different locations may experience varying illuminations even in the same scene. (b) The appearance of the same target may change significantly owing to different illuminations. (c) Our target highlighting scheme for image-based and video-based inputs.

**Target highlighting.** Highlighting the target is related to both the target itself and the nature of the task. In tasks such as object recognition or detection relying on static images, attention should be paid to the location of the target region. For others like action recognition or tracking, the emphasis moves toward capturing the movement trajectory of the target. Accordingly, we leverage saliency for static highlighting and TGMM for dynamic attention as depicted in Fig.3(c).

**Saliency for static highlighting.** For the semantic vision tasks, the target should be in the most attractive region corresponding to the human visual attention mechanism. According to [16], the salient objects should have higher spectral residuals. Therefore, we utilize the spectral residual to produce a saliency map. The spectral residual can be calculated as $R(I) = L(I) - L(I) * H$, where $R(I)$ denotes the spectral residual of input image $I$, representing the statistical singularities that are particular to $I$ in the frequency domain. $L(I)$ represents the log amplitude spectrum, computed as $L(I) = \log(|\mathcal{F}(I)|)$, where $\mathcal{F}(\cdot)$ is the Fast Fourier Transform (FFT) of $I$. $H$ is the averaged spectrum that can be approximated by convolution. It highlights areas of the image that differ from the norm and thus are more likely to attract visual attention. Then we need to map it back to the spatial domain using the Inverse Fast Fourier Transform (IFFT), and we have the saliency map as $S(I) = \mathcal{F}^{-1}\{\mathrm{e}^{R(I)+i\cdot P(\mathcal{F}(I))}\}$, where $S(I)$ represents the saliency map obtained after mapping back to the spatial domain, $\mathcal{F}^{-1}$ denotes IFFT processing, and $P(\mathcal{F}(I))$ is the phase spectrum of the image. Extracting the salient region can derive the static attention map.

**TGMM for dynamic highlighting.** The video-based tasks are strongly correlated with temporal information. Therefore, the temporal changes of the target should be paid more attention to when implementing target-aware masking strategies for video inputs. To this end, we utilize TGMM to learn the information of dynamic targets. TGMM expresses pixels in a set of Gaussian distributions, capturing the variations of pixel value along with time. It can be formulated as:

$$B(x) = \sum_{i=1}^{K} \omega_{i,t}(x) \cdot \mathcal{N}(I_t(x); \mu_{i,t}(x), \sigma_{i,t}(x)), \tag{3}$$

where $B(x)$ represents the distribution map of TGMM, $\mathcal{N}$ denotes a Gaussian distribution, and $I_t(x)$ is the observed value of pixel $x$ at time $t$. $\omega_{i,t}(x)$ is the weight of the $i$-th Gaussian distribution at time $t$, with $\mu_{i,t}(x)$ and $\sigma_{i,t}(x)$ representing the mean and covariance matrix of the $i$-th Gaussian distribution, where $i$ ranges from 1 to $K$. The movement of the target entity will lead to a mismatch between the pixel values within the corresponding region and the Gaussian distribution of the background model. When this mismatch becomes sufficiently distinct, it indicates the presence of the target. Then a dynamic attention map can be created through binary processing, highlighting areas of significant change or movement.

**Region-dependent masking strategy.** Masking is often used in SSL since it helps to learn critical features and improve the generalization ability of the model. In our scheme, we also adopt a target region-dependent masking strategy. Following [13], we first divide the image into patches. Obviously, they fall into target and non-target regions. For patches within the target region, we apply an erasing strategy, where the content will be removed for a robust representation as per [51]. For background regions, we employ a leaky-patching strategy, which is inspired by the "Leaky-ReLU" activation function and "leaks" background information in the non-target regions to gain diversity in representation. The mask ratio and size of the erasing mask are smaller than those of the leaky-present mask to preserve the effective target-relevant information.

**Robust target feature learning.** With the target-aware views achieved by masking, we investigate how to use them for illumination-robust feature representations toward semantic vision tasks. Similarly, to mitigate divergence issues caused by negative samples, we employ a non-negative contrastive learning loss from BYOL [10] to achieve it:

$$\mathcal{L}_{\text{BYOL}}\left(v, v^{+}\right)=2-\frac{2 \cdot\left\langle z(p(F(v))), p^{\prime}\left(F^{\prime}\left(v^{+}\right)\right)\right\rangle}{\|z(p(F(v)))\|_{2} \cdot\left\|p^{\prime}\left(F^{\prime}\left(v^{+}\right)\right)\right\|_{2}}, \tag{4}$$

where $v$ and $v^+$ are the anchor and positive sample, respectively. $z$ is the prediction head. $p/p'$ are the projection heads, and $F/F'$ are the feature extractors. They share the same structure but the gradient is only passed in $F$ and $p$, while their counterparts are updated by Exponential Moving Average (EMA). In our feature representation learning scheme, we have two branches: one is the original model trained on daytime data, and the other is fed with $I_m$ obtained via the domain alignment. As only the anchor branch has a gradient, we employ a dual-branch optimization strategy to obtain the optimal feature representation. Then the final contrastive learning loss can be written as:

$$\mathcal{L}_{CL}=\mathcal{L}_{\text{BYOL}}(I_b, I_m)+\mathcal{L}_{\text{BYOL}}(I_m, I_b). \tag{5}$$

### 3.4 Network Training

The overall structure and losses for training are illustrated in Fig.2. During the training phase, the low-lit data is only used in the image-level domain alignment stage, and it is not coupled with what is used for the well-lit branch, which plays a more important role in refining the feature representation. For the inference stage, the network is finalized and no well-lit data is required. The low-lit one is set as the sole input to yield the result for high-level vision tasks. When training for the domain alignment, the image-level adversarial loss and feature-level MMD loss are utilized, then the entire optimization objective can be expressed as:

$$\mathcal{L}_{align}=\mathcal{L}_{adv}+\alpha\mathcal{L}_{MMD}, \tag{6}$$

where $\alpha$ is the balancing parameter, of which the value varies along with different tasks. After that, the target-aware representation learning is achieved together with the downstream task. For this stage, the loss can be expressed as:

$$\mathcal{L}_{rep}=\mathcal{L}_{task}^{b}+\mathcal{L}_{task}^{i}+\beta\mathcal{L}_{CL}, \tag{7}$$

where $\beta$ is also a task-dependent balancing parameter. $\mathcal{L}_{task}$ is the loss for downstream task. For example, it will be the cross-entropy loss for recognition tasks, and bounding box regression loss for detection tasks. The superscripts $b$ and $i$ denote the branches of the original model trained on daytime data and the intermediate state input $I_m$, respectively.

# 4 Experiments

## 4.1 Experimental Setup

The proposed method is implemented with PyTorch on an NVIDIA RTX 3090 GPU. As our method attempts to construct a general framework for high-level vision tasks, we select several typical tasks for validation, including 1)image-based object recognition (classification), 2)image-based object detection, and 3) video-based action recognition. We evaluate our method on multiple datasets. The object recognition task is verified on two datasets of CODaN [22] and ExDark [33]. The ExDark dataset is also employed for object detection. For the more challenging action recognition task, we use the ARID dataset [54], which is also the benchmark for UG2+ 2022 Action Recognition from Dark Video Challenge. More details about those datasets and the network parameter settings can be found in the Supplementary file.

## 4.2 Object Recognition

**Result on CODaN.** From Table 1, it is evident that low-light enhancement methods, which enhance images pixel-wise from the perspective of human visual perception, underperform in downstream vision tasks. Meanwhile, the performance differences among various enhancement methods are not significant, with a maximum variation of about 2%. It indicates that the gap between low-level and high-level vision tasks is not attributable to specific methods but is a common phenomenon across the board. In contrast, methods based on domain generalization and adaptation show a greater divergence in performance. The accuracy of MixStyle [60] is only 53.12%, which is even lower than that of enhancement methods. It may be due to MixStyle's reliance on low-level style features like color and shape. Meanwhile, Similarity Min-Max [34] achieves a much better result of 65.87%, likely because it implements dual-level adaptation at both the image and feature levels. In comparison, our method has achieved an improvement of about 5% on this basis. This proves that bi-directional alignment can effectively reduce complexity, and focusing on the target area rather than the entire image holds significant value for high-level visual tasks.

**Result on ExDark.** The classification results on the ExDark dataset [33] are demonstrated in Table 2. Although this dataset is originally intended for detection purposes, it is also widely used for verifying the day-night adaptation performance since its data is also annotated with class labels. Compared with CODaN, this dataset can be more challenging since the target is smaller and the illumination condition is worse. According to the description in [33], we extract 400 normal light images per class from the COCO dataset as training data, totaling 4,800 images, and evaluate on the complete ExDark dataset. As can be seen, when facing darker conditions and smaller subject targets in ExDark dataset, the limitations of low-light enhancement methods become even more apparent. RUAS [32] leads to many overexposed areas and destroys the structure of the images, making even worse results than the baseline without any enhancing processes. For domain adaptation-based methods, IRM [1] and CIConv [22] rely on detecting invariant low-level features like edge and color also fail since there is a significant loss of texture information in extremely dark environments. By contrast, our framework can still witness about 2% improvement when compared with the second best one of Similarity Min-Max [34]. It proves the bi-directional alignment can alleviate the information loss owing to darkness and focusing on the target can make the feature more robust to extreme illumination conditions.

**Table 1:** Object recognition performance on CODaN dataset.

| **Method** | Accuracy(%) |
|---|---|
| ResNet-18 [15](baseline) | 53.32 |
| **Low-Light Enhancement** | |
| EnlightenGAN [20] | 56.68 |
| LEDNet [61] | 57.40 |
| Zero-DCE++ [23] | 57.96 |
| RUAS [32] | 58.36 |
| SCI [35] | 58.68 |
| URetinexNet [52] | 58.72 |
| **Domain Generalization & Adaption** | |
| MixStyle [60] | 53.12 |
| IRM [1] | 54.52 |
| AdaBN [25] | 55.55 |
| MAET [8] | 56.48 |
| CIConv [22] | 60.32 |
| Similarity Min-Max [34] | 65.87 |
| **Ours** | **70.32** |

**Table 2:** Object recognition performance on ExDark dataset.

| **Method** | Accuracy(%) |
|---|---|
| ResNet-50 [15](baseline) | 43.15 |
| **Low-Light Enhancement** | |
| EnlightenGAN [20] | 44.63 |
| LEDNet [61] | 43.90 |
| Zero-DCE++ [23] | 45.99 |
| RUAS [32] | 42.18 |
| SCI [35] | 46.45 |
| URetinexNet [52] | 46.37 |
| **Domain Generalization & Adaption** | |
| MixStyle [60] | 46.46 |
| IRM [1] | 44.64 |
| AdaBN [25] | 46.84 |
| MAET [8] | 46.04 |
| CIConv [22] | 45.17 |
| Similarity Min-Max [34] | 49.03 |
| **Ours** | **51.58** |

### 4.3 Object Detection

In Table 3, we also validate the performance of dark object detection. Following pioneer works [8, 33], to match the normal light data with the 12 categories in ExDark dataset, we randomly selected 1000 images for each category from the COCO and VOC datasets, totaling 12,000 images, as the well-lit training data, and the validation set of ExDark is used as the low-lit test set. As observed, this time all the low-lit enhancement methods witness a failure compared with the bare baseline. It implies the noise introduced during the enhancement may be more harmful to the localization task. The recent domain adaptation methods [34] can improve the performance, but the severe noise in ExDark makes localization challenging. Despite these challenges, our method still achieves the best performance in the object detection task, with the mAP exceeding the second-best one [34] by about 3.4.

### 4.4 Action Recognition

Following the image-based tasks, we explore another challenging task of video-based action recognition, which is studied in the UG2+ 2022 Action Recognition from Dark Video Challenge. We also evaluate the network as the unsupervised learning track in this challenge. As illustrated in Table 4, in more complex low-light action recognition tasks, continuous changes in illumination over time can result in more noise if temporal consistency is not maintained globally during enhancement. This may disrupt the coherence of features across temporal variations and result in even worse performance. Thanks to the target highlighting strategy, our feature presentation learning pays more attention to the motion of targets. As a result, the learned feature is more robust to changes in illumination and backgrounds. It leads to a more substantial improvement over Luo *et al.*'s [34] at about 6% compared with image-based recognition tasks. It is also worth mentioning that although the fully supervised SOTA is very high, our unsupervised method narrows the gap between unsupervised and supervised learning by about 10 % with a smaller backbone than [15, 45]. It even surpasses some fully supervised models like C3D [44] and 3D-ResNet-18 [15], validating the effectiveness of our feature representation learning scheme.

**Table 3:** Object Detection performance on Exdark dataset.

| **Method** | mAP@0.5 |
|---|---|
| YOLOv3 [38](baseline) | 58.7 |
| YOLOv3-fully supervised | 76.4 |
| **Low-Light Enhancement** | |
| EnlightenGAN [20] | 51.4 |
| Zero-DCE [11] | 57.1 |
| URetinexNet [52] | 56.6 |
| LEDNet [61] | 51.4 |
| RUAS [32] | 56.3 |
| SCI [35] | 56.9 |
| IAT-YOLO [7] | 55.9 |
| **Domain Generalization & Adaption** | |
| MAET [8] | 57.3 |
| HLA-face [50] | 57.6 |
| Similarity Min-Max [34] | 60.8 |
| **Ours** | **64.2** |

**Table 4:** Action Recognition performance on ARID dataset

| **Method** | Accuracy(%) |
|---|---|
| I3D [4] (baseline) | 47.02% |
| **Fully Supervised** | |
| C3D [44] | 46.17% |
| 3D-ResNet-18 [15] | 54.68% |
| I3D (fully supervised ver.) [4] | 68.29% |
| 3D-ResNet-101 [15] | 71.57% |
| DTCM [45] | 96.36% |
| **Low-Light Enhancement** | |
| StableLLVE [58] | 45.08% |
| SMOID [19] | 47.27% |
| SGZ [59] | 46.42% |
| **Domain Generalization & Adaption** | |
| AdaBN [25] | 46.17% |
| Similarity Min-Max [34] | 52.01% |
| **Ours** | **58.08%** |

### 4.5 Ablation Studies

In this section, we first conduct a detailed experiment on each component of the framework to deeply study the effectiveness of our designs, including the domain alignment scheme and target-aware masked representation learning. Then we also test different domain-transferring strategies to demonstrate why our design can be more useful for high-level tasks. For simplicity, we choose CODaN as the benchmark. Without specific saying, the experimental settings are the same as what is mentioned in Section 4.1. More experiment results can be found in the Supplementary file.

**Impacts of components of the proposed framework.** In Table 5, we show how our framework helps to learn a robust feature representation for low-illumination high-level vision tasks. Compared with the baseline (with all modules marked "×") of ResNet-18, the domain alignment makes the dark meet normal light, resulting in a significant improvement at about 8%. On the other hand, the target-aware representation learning scheme seems more important since it makes a more significant improvement at about 10%. We are also surprised that even without a SSL paradigm for optimizing the feature representation, the target-aware masking can even achieve competitive performance. It emphasizes the role of a good view of the target in high-level tasks.

**Table 5:** Impacts of components of the proposed framework.

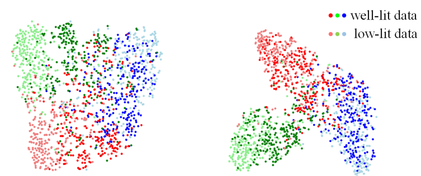
| two-level bi-directional alignment | | target-aware mask representation learning | | Accuracy(%) |
|---|---|---|---|---|
| image-level | feature-level | target-aware mask | contrastive learning | |
| × | × | × | × | 53.32 |
| √ | × | × | × | 59.08 |
| √ | √ | × | × | 61.64 |
| √ | √ | √ | × | 64.40 |
| √ | √ | × | √ | 64.32 |
| √ | √ | √ | √ | 70.32 |

**Effect of different domain transferring strategies.** As shown in Table 6, the performance of bright-to-dark unidirectional adaptation is much lower than that of dark-to-bright. It suggests enhanced information, even with noise, can help to yield better results than removing it. Compared to it, our image-level bi-directional alignment can make a further 2% improvement. Moreover, the feature-level alignment achieves better performance, which may be because this framework is oriented toward high-level tasks.

**Visualization.** We also present the visualization results of the ResNet-18 backbone and our framework, encoded using t-SNE. Fig.4 illustrates the feature distribution across different illumination conditions. The different colors of red, green, and blue refer to different classes, while the intensity of the color indicates whether the data is in low or sufficient light. It is apparent that our framework narrows the distance between the features of different illuminations for one class, and correspondingly the inter-class distance is enlarged, resulting in a more robust illumination-independent feature and enhancing overall performance.

**Table 6:** Results of different domain transferring strategies

| Strategies | Accuracy(%) |
|---|---|
| **unidirectional domain adaptation** | |
| dark→bright | 57.96 |
| bright→dark | 51.52 |
| **bi-directional domain alignment** | |
| image-level only | 59.08 |
| feature-level only | 60.56 |
| **Ours (bi-direct. align.)** | **61.64** |




(a) original daytime model (b) ours

**Fig. 4:** Visualizing the projection of feature distribution of inter-illumination conditions of the baseline and our framework.

## 5 Conclusion

In this study, we address the challenge of low illumination for high-level vision tasks via a target-aware representation learning framework. To address the gap between low-lit and well-lit domains, we first propose a two-level bi-directional domain alignment, considering both image appearance and semantic features, to bridge the gap between data collected under varying illumination conditions. To harness the high-level features of the target, we introduce a frequency-based saliency mechanism and a Temporal Gaussian Mixture Model-based approach to highlight task-relevant elements from static and dynamic views, respectively. Additionally, we devise a mask token-based representation learning scheme for learning robust target-aware features. Extensive experiments on CODaN, ExDark, and ARID datasets validate the effectiveness of our method in enhancing feature representation for both image- and video-based high-level vision tasks, including classification, detection, and action recognition.

## Acknowledgements

The work is jointly supported by the National Natural Science Foundation of China under grants No.62002271, and 62272364, the National Science and Technology Major Project under grant No.2022ZD0117103, the provincial Key Research and Development Program of Shaanxi under grant No.2024GH-ZDXM-47, the Research Project on Higher Education Teaching Reform of Shaanxi Province under grant No.23JG003.

## References

1. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)

2. Arruda, V.F., Paixao, T.M., Berriel, R.F., De Souza, A.F., Badue, C., Sebe, N., Oliveira-Santos, T.: Cross-domain car detection using unsupervised image-to-image translation: From day to night. In: IJCNN. pp. 1–8 (2019)
3. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In: ICCV. pp. 12504–12513 (October 2023)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
6. Chen, W., Wenjing, W., Wenhan, Y., Jiaying, L.: Deep retinex decomposition for low-light enhancement. In: BMVC. pp. 1–12 (2018)
7. Cui, Z., Li, K., Gu, L., Su, S., Gao, P., Jiang, Z., Qiao, Y., Harada, T.: You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. In: BMVC. pp. 21–24 (2022)
8. Cui, Z., Qi, G.J., Gu, L., You, S., Zhang, Z., Harada, T.: Multitask aet with orthogonal tangent regularity for dark object detection. In: ICCV. pp. 2553–2562 (2021)
9. Feichtenhofer, C., Li, Y., He, K., et al.: Masked autoencoders as spatiotemporal learners. In: NeurIPS. pp. 35946–35958 (2022)
10. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: NeurIPS. pp. 21271–21284 (2020)
11. 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: CVPR. pp. 1780–1789 (2020)
12. Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. IEEE TIP **26**(2), 982–993 (2016)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
16. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: CVPR. pp. 1–8 (2007)
17. Huang, S., Tran, T.D.: Sparse signal recovery via generalized entropy functions minimization. IEEE TSP **67**(5), 1322–1337 (2018)
18. Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.M., Fu, D., Shen, X., Feng, J.: Contrastive masked autoencoders are stronger vision learners. IEEE TPAMI **46**(4), 2506–2517 (2023)
19. Jiang, H., Zheng, Y.: Learning to see moving objects in the dark. In: ICCV. pp. 7324–7333 (2019)
20. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE TIP **30**, 2340–2349 (2021)
21. Land, E.H.: The retinex theory of color vision. Scientific American **237**(6), 108–129 (1977)
22. Lengyel, A., Garg, S., Milford, M., van Gemert, J.C.: Zero-shot day-night domain adaptation with a physics prior. In: ICCV. pp. 4399–4409 (2021)

23. Li, C., Guo, C., Loy, C.C.: Learning to enhance low-light image via zero-reference deep curve estimation. IEEE TPAMI **44**(8), 4225–4238 (2021)
24. Li, M., Liu, J., Yang, W., Sun, X., Guo, Z.: Structure-revealing low-light image enhancement via robust retinex model. IEEE TIP **27**(6), 2828–2841 (2018)
25. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive batch normalization for practical domain adaptation. PR **80**, 109–117 (2018)
26. Li, Y., Chen, H., Feng, G., Miao, Q.: Learning robust representations with information bottleneck and memory network for rgb-d-based gesture recognition. In: ICCV. pp. 20968–20978 (2023)
27. Li, Y., Chen, H., Miao, Q., Ge, D., Liang, S., Ma, Z., Zhao, B.: Image hazing and dehazing: From the viewpoint of two-way image translation with a weakly supervised framework. IEEE TMM **25**, 4704–4717 (2023)
28. Li, Y., Miao, Q., Ouyang, W., Ma, Z., Fang, H., Dong, C., Quan, Y.: Lap-net: Level-aware progressive network for image dehazing. In: ICCV. pp. 3276–3285 (2019)
29. Li, Y., Qi, T., Ma, Z., Quan, D., Miao, Q.: Seeking a hierarchical prototype for multimodal gesture recognition. IEEE TNNLS (2023)
30. Liang, Z., Li, C., Zhou, S., Feng, R., Loy, C.C.: Iterative prompt learning for unsupervised backlit image enhancement. In: ICCV. pp. 8094–8103 (2023)
31. Liu, D., Wen, B., Liu, X., Wang, Z., Huang, T.: When image denoising meets high-level vision tasks: A deep learning approach. In: IJCAI. pp. 842–848 (2018)
32. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: CVPR. pp. 10561–10570 (2021)
33. Loh, Y.P., Chan, C.S.: Getting to know low-light images with the exclusively dark dataset. CVIU **178**, 30–42 (2019)
34. Luo, R., Wang, W., Yang, W., Liu, J.: Similarity min-max: Zero-shot day-night domain adaptation. In: ICCV. pp. 8104–8114 (2023)
35. Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z.: Toward fast, flexible, and robust low-light image enhancement. In: CVPR. pp. 5637–5646 (2022)
36. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV. pp. 319–345 (2020)
37. Porav, H., Maddern, W., Newman, P.: Adversarial training for adverse conditions: Robust metric localisation using appearance transfer. In: ICRA. pp. 1011–1018 (2018)
38. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
39. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: ICCV. pp. 7374–7383 (2019)
40. Sasagawa, Y., Nagahara, H.: Yolo in the dark-domain adaptation method for merging multiple models. In: ECCV. pp. 345–359 (2020)
41. Shi, Y., Fu, X., Li, Y., Miao, K., Liu, X., Zhao, B., Miao, Q.: A semi-supervised underexposed image enhancement network with supervised context attention and multi-exposure fusion. IEEE TMM **26**, 1229–1243 (2023)
42. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: NeurIPS. pp. 10078–10093 (2022)
43. Trahanias, P.E., Venetsanopoulos, A.N.: Color image enhancement through 3-d histogram equalization. In: ICPR. pp. 545–548 (1992)
44. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)

45. Tu, Z., Liu, Y., Zhang, Y., Mu, Q., Yuan, J.: Dtcm: Joint optimization of dark enhancement and action recognition in videos. IEEE TIP **32**, 3507–3520 (2023)
46. VidalMata, R.G., Banerjee, S., RichardWebster, B., Albright, M., Davalos, P., McCloskey, S., Miller, B., Tambo, A., Ghosh, S., Nagesh, S., et al.: Bridging the gap between computational photography and visual recognition. IEEE TPAMI **43**(12), 4272–4290 (2020)
47. Wang, L., Ouyang, W., Wang, X., Lu, H.: Visual tracking with fully convolutional networks. In: ICCV. pp. 3119–3127 (2015)
48. Wang, S., Zheng, J., Hu, H.M., Li, B.: Naturalness preserved enhancement algorithm for non-uniform illumination images. IEEE TIP **22**(9), 3538–3548 (2013)
49. Wang, W., Xu, Z., Huang, H., Liu, J.: Self-aligned concave curve: Illumination enhancement for unsupervised adaptation. In: ACM MM. pp. 2617–2626 (2022)
50. Wang, W., Yang, W., Liu, J.: Hla-face: Joint high-low adaptation for low light face detection. In: CVPR. pp. 16195–16204 (2021)
51. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: CVPR. pp. 1568–1576 (2017)
52. 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: CVPR. pp. 5901–5910 (2022)
53. Wu, X., Wu, Z., Guo, H., Ju, L., Wang, S.: Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In: CVPR. pp. 15769–15778 (2021)
54. Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J., See, S.: Arid: A new dataset for recognizing action in the dark. In: IJCAI Workshop. pp. 70–84 (2021)
55. Yang, W., Tan, R.T., Wang, S., Fang, Y., Liu, J.: Single image deraining: From model-based to data-driven and beyond. IEEE TPAMI **43**(11), 4059–4077 (2020)
56. Yang, W., Yuan, Y., Ren, W., Liu, J., Scheirer, W.J., Wang, Z., Zhang, T., Zhong, Q., Xie, D., Pu, S., et al.: Advancing image understanding in poor visibility environments: A collective benchmark study. IEEE TIP **29**, 5737–5752 (2020)
57. Yao, Y., Desai, N., Palaniswami, M.: Masked contrastive representation learning. arXiv preprint arXiv:2211.06012 (2022)
58. Zhang, F., Li, Y., You, S., Fu, Y.: Learning temporal consistency for low light video enhancement from single images. In: CVPR. pp. 4967–4976 (2021)
59. Zheng, S., Gupta, G.: Semantic-guided zero-shot learning for low-light image/video enhancement. In: WACV. pp. 581–590 (2022)
60. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: ICLR. pp. 1–15 (2021)
61. Zhou, S., Li, C., Change Loy, C.: Lednet: Joint low-light enhancement and deblurring in the dark. In: ECCV. pp. 573–589 (2022)
62. Zhu, J., Xia, Y., Wu, L., Deng, J., Zhou, W., Qin, T., Liu, T.Y., Li, H.: Masked contrastive representation learning for reinforcement learning. IEEE TPAMI **45**(3), 3421–3433 (2022)