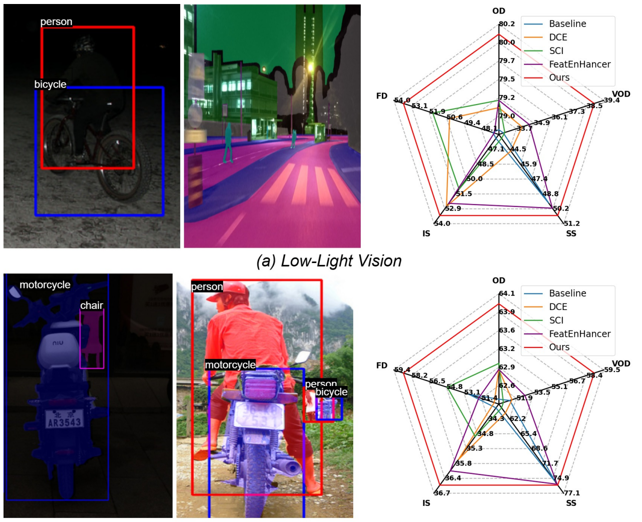
TorchAdapt: Towards Light-Agnostic Real-Time Visual Perception

Khurram Azeem Hashmi[1,2] Karthik Palyakere Suresh[1,2] Didier Stricker[1,2]
Muhammad Zeshan Afzal[1,2]
[1]DFKI [2]RPTU Kaiserslautern-Landau
khurram_azeem.hashmi@dfki.de

Abstract

Low-light conditions significantly degrade the performance of high-level vision tasks. Existing approaches either enhance low-light images without considering normal illumination scenarios, leading to poor generalization, or are tailored to specific tasks. We propose ***TorchAdapt****, a real-time adaptive feature enhancement framework that generalizes robustly across varying illumination conditions without degrading performance in well-lit scenarios. TorchAdapt consists of two complementary modules: the* ***Torch*** *module enhances semantic features beneficial for downstream tasks, while the* ***Adapt*** *module dynamically modulates these enhancements based on input content. Leveraging a novel light-agnostic learning strategy, TorchAdapt aligns feature representations of enhanced and well-lit images to produce powerful illumination-invariant features. Extensive experiments on multiple high-level vision tasks, including object detection, face detection, instance segmentation, semantic segmentation, and video object detection, demonstrate that TorchAdapt consistently outperforms state-of-the-art low-light enhancement and task-specific methods in both low-light and light-agnostic settings. TorchAdapt thus provides a unified, flexible solution for robust visual perception across diverse lighting conditions.*



(a) Low-Light Vision

(b) Light-Agnostic Vision

Figure 1. Prior downstream methods in low-light vision assume that all samples belong to low-light conditions as shown in **(a)**, leading to underperformance in varying illumination environments, depicted in **(b)**. **TorchAdapt** alleviates this bottleneck via learning powerful illumination-invariant features and achieves consistent gains across low-light and light-agnostic settings on Object Detection (OD), Face Detection (FD), Instance Segmentation (IS), Semantic Segmentation (SS), and Video Object Detection (VOD). Best viewed on the screen.

1. Introduction

Recent advances in visual perception tasks [13, 31, 77, 82] are highly effective at extracting features for complex tasks when provided with high-quality inputs. However, replicating this performance is challenging when data is scarce, noisy, or expensive to obtain. A common issue is varying illumination, particularly low-light conditions, which hinders the robustness and reliability of safety-critical visual perception systems [1, 3, 7, 63].

A common strategy to improve low-light performance involves pre-processing images with Low-Light Image Enhancement (LLIE) methods [17, 24, 33, 44, 50, 51, 71, 84, 85] before feeding them into high-level vision models. However, this simple integration often leads to weaker gains [16, 26], as these arts conform to human visual perception (optimized on pixel-level objectives like L1 or MSE losses [4]) instead of machine visual perception. Moreover, applying these enhancement methods indiscriminately can degrade the performance on images/videos captured under standard lighting conditions, as alterations intended for low-light scenarios [47, 60, 80, 81] may be unnecessary or even detrimental in well-lit environments [15, 40, 58, 78].

To address low-light vision tasks, recent methods employ transfer learning [16, 59, 72] and domain adaptation techniques [20, 32, 48]. Despite improvements, these efforts are carefully designed for a single high-level vision task, such as object detection [16] or semantic segmentation [59, 72], and require cumbersome multi-stage training to adapt to each dense task [20, 32, 48]. Moreover, they often assume consistently low-light data, leading to performance degradation on well-lit images and videos, sometimes worse than baselines (see Table 2). In practical applications with unpredictable lighting—such as autonomous driving [7, 37] at different times of the day, surveillance systems [63] monitoring both indoor and outdoor environments or robots [3] operating between well-lit and low-lit areas—these limitations become critical. This highlights the need for high-level vision models to be **light-agnostic**—*capable of maintaining high performance across varying lighting conditions*, as visualized in Fig. 1(b).

In this work, we propose **TorchAdapt**, a real-time, light-agnostic adaptive feature enhancement framework designed to improve the performance of any high-level vision task under varying illumination. TorchAdapt comprises two primary modules: *Torch* and *Adapt*, illustrated in Fig. 2. The *Torch* module is a lightweight feature enhancement network that enriches semantic representations favorable for downstream vision tasks. The *Adapt* module learns to modulate this feature enhancement dynamically based on input data, amplifying or reducing the effect of the *Torch* module.

TorchAdapt builds on the consensus [26, 38] that low-to-normal light adaptation lacks a physical model, leading to unexpected artifacts during optimization when transforming images in pixel space. Therefore, aligning representations in the feature space is a more practical approach for high-level vision tasks [48, 65]. We optimize TorchAdapt using a non-contrastive self-supervised learning objective [12, 23] to align representations from low-light and well-lit images. However, unlike prior works operating in a similar direction [16, 48, 65], *TorchAdapt also learns not to adversely affect representations of well-lit images*. As illustrated in Fig. 3, we take a well-lit image V_1 and its synthetically generated low-light counterpart V_2, pass them through TorchAdapt to obtain enhanced images $\hat{V}_1$ and $\hat{V}_2$, and feed them along with V_1 into a pre-trained frozen visual encoder [19, 27, 55] to generate representations. Since the encoder is frozen, only TorchAdapt is optimized to help the encoder produce well-lit representations (like V_1) for $\hat{V}_1$ and $\hat{V}_2$.

Thanks to the conceptual simplicity, TorchAdapt enjoys two benefits. ❶ It can be trained with a simple self-supervised loss on a general-purpose dataset like ImageNet [18], circumventing large-scale real-world low-light datasets. ❷ The training of TorchAdapt is independent of the high-level vision task and needs to be done only once to align with the backbone networks [6, 19, 27]. Later, TorchAdapt can be plugged into any downstream tasks in a light-agnostic manner. We demonstrate the effectiveness of TorchAdapt on five representative visual perception tasks, including object detection [21, 47], face detection [78, 80], semantic segmentation [15, 60], instance segmentation [10, 40], and video object detection [58, 81] under both low-light and light-agnostic settings. For all tasks, TorchAdapt brings consistent and significant gains over baselines under both low-light and light-agnostic settings, outperforming previous state-of-the-art task-specific and LLIE methods (see Fig. 1). Moreover, TorchAdapt accomplishes all this with negligible computational overhead (see Table 6).

Contributions. *Method:* We propose TorchAdapt, a real-time, light-agnostic adaptive feature enhancement framework that improves the performance of any high-level vision task under varying illumination without compromising accuracy in well-lit conditions. ***Simplicity:*** Thanks to its architectural innovations and self-supervised objective, TorchAdapt significantly simplifies the training strategy for any high-level vision task pipeline, circumventing corresponding paired images/videos, as shown in Fig 3. ***Flexibility:*** TorchAdapt is independent of vision encoders and downstream task architectures, demonstrating model-agnostic generality. (see § 5). Any zero-reference LLIE method [24, 33, 50, 67] can serve as the *Torch* module in our framework to achieve light-agnostic performance (see Fig. 5). ***Results:*** We conduct extensive experiments on five visual perception tasks involving images and videos under both low-light and light-agnostic settings using curated datasets. TorchAdapt outperforms state-of-the-art task-specific and LLIE methods across all settings when integrated with the same baselines (see Fig. 1).

2. Related Work

2.1. Enhancing Low-Light Images

Low-light Image Enhancement (LLIE) aims to improve the visual quality of images captured under insufficient lighting conditions. Modern learning-based approaches have become prevalent, thanks to the emergence of low-light datasets [5, 8, 25, 43, 70]. Supervised methods [70, 73, 74, 84, 86] leverage these datasets to design effective networks trained to predict normal-light images from low-light ones. Unsupervised LLIE approaches [30, 34, 79] employ adversarial learning from unpaired supervision to relax the requirement of paired normal-low light data. However, all of these works still rely on specific training data, which limits their generalizability to unseen scenarios. Another line of work known as zero-reference methods [24, 33, 44, 50, 67] get rid of both paired and unpaired data to enhance low-light images. These arts design enhancement networks that optimize on a set of non-reference loss functions. Since these works are mainly designed to conform to human visual perception, they bring sub-optimal gains for machine visual perception tasks. Furthermore, they are not light-agnostic (capable of handling both normal and low-light images). This work bridges this gap by complementing these zero-reference methods,

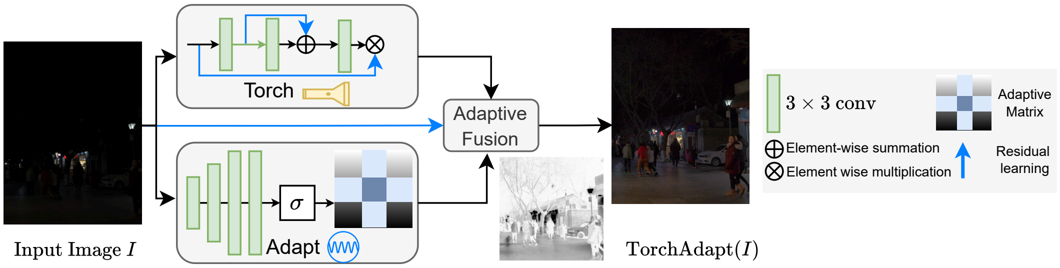


Figure 2. **TorchAdapt architectural overview.** The *Torch* module enhances feature representation by applying 3×3 convolutions and element-wise operations (Eq. 1). The *Adapt* module performs adaptive modulation, creating a content-aware response based on the input I (Eq. 2). Both modules are fused via *Adaptive Fusion*, integrating feature enhancements from Torch and contextual adaptations from Adapt. This adaptive fusion produces illumination-invariant features, improving robustness across varying lighting conditions. Adaptive Fusion is explained in Eq. 4.

where they can be employed as a *Torch* in our TorchAdapt framework to improve light-agnostic visual perception.

2.2. Low-Light Visual Perception

This direction of work considers machine perception as the criteria for success while enhancing low-light images.

Low-light Object Detection (LLOD). Object detection, as a fundamental problem in computer vision, has also been well-explored in low-light vision [16, 20, 39, 51, 61, 62, 64–66]. Based on the dataset, objects vary and can be generic [47, 52] or specific like pedestrians [53], and human faces [80]. One category of LLOD methods [51, 56, 62, 65] jointly learns the enhancement and detection task in an end-to-end fashion to improve the overall detection performance. Other approaches involve learning a low-light detector through fusing multiple models [61] or resorting to domain adaptation frameworks [16, 20, 64], minimizing the distribution variance between normal and low-light images. Most of these works are specifically designed to tackle the dark object detection task. Furthermore, when these methods are optimized for LLOD, they underperform on normal-light images. In contrast, TorchAdapt is fully task-agnostic for any high-level vision task, and it does not worsen the performance of the detector on normal-light data.

Low-light Tasks-Agnostic Methods. Recently, various low-light task-agnostic methods [20, 26, 32, 48, 65] have emerged in the literature. Unlike prior approaches, they are more flexible and can be integrated to improve any downstream task under low-light vision. Our work follows the same spirit and offers similar flexibility. FeatEnHancer [26], as one of the representative works, learns to enhance the hierarchical features suitable for downstream tasks. Other seminal attempts [20, 32, 48] treat low-light perception tasks as domain adaption (DA) learning. They exploit well-lit source domain data [40, 78] and learn to adapt to low-light target domain data [47, 80]. Although flexible, the DA methods rely on multi-stage training frameworks for each task [48]. In contrast, our TorchAdapt is pre-trained with the frozen backbone on the domain-independent data like ImageNet [18]. Moreover, it is important to recognize the efforts of recent real low-light datasets [10, 35, 60, 68, 75, 81] that enable the increasing research on several downstream tasks in low-light vision.

2.3. Learning Equivariant Representations in Low-Light Vision with Self-Supervised Objectives

Annotating low-light vision datasets is a much more difficult task than annotating well-lit datasets. Hence, it is a common practice to apply self-supervised learning to learn illumination-invariant features [16, 65] or aligning well-lit and low-light data distributions [38, 48, 66]. MAET [16] adopts the autoencoding transformation [83] to decode illumination-degrading parameters and detection predictions. SACC [65] merges the pretext tasks of rotation prediction [22] and jigsaw puzzling [54] to learn illumination enhancement from low-light images. Most related to our work is SimMinMax [48], which adopts the BYOL [23] framework to learn model-level adaptation by maximizing the similarity of features between images at night and daytime. However, since SimMinMax [48] learns to adapt from daytime to nighttime images, it needs to be trained for each high-level vision task, requiring the corresponding daytime labeled dataset. In contrast, TorchAdapt leverages self-supervised objectives to align backbone features during the pretext task. Furthermore, it can be trained on any general-purpose dataset. This significantly simplifies the training procedure of TorchAdapt, as compared to SimMinMax, on any high-level vision task. TorchAdapt is also unique to prior works in its ability to handle light-agnostic data.

3. TorchAdapt

This section formally introduces TorchAdapt, explaining the proposed **Torch** in § 3.1 and **Adapt** in § 3.2 modules, followed by the light-agnostic learning in § 3.3.

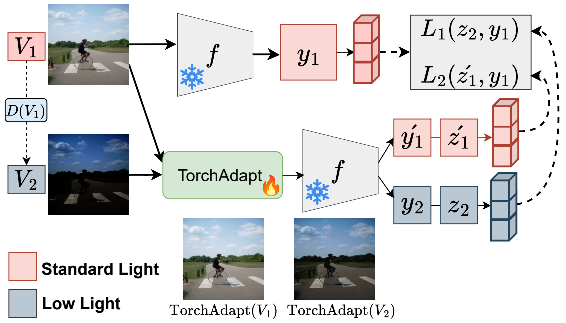


Figure 3. **Light-Agnostic learning pipeline.** $D(.)$ is a darkening function applied to the well-lit image V_1 to synthesize the low-light view V_2. Both V_1 and V_2 are processed by TorchAdapt and then fed into a frozen pre-trained vision encoder f. A cosine similarity loss optimizes TorchAdapt to produce well-lit representations for both V_1 and V_2, enabling it to learn illumination-invariant features. The $TorchAdapt(.)$ function integrates Eq. 1 and Eq. 4. Refer to § 3.3 for a complete explanation of each symbol.

3.1. Torch

The Torch module in our TorchAdapt framework serves as an enhancement network as in Low-light Image Enhancement (LLIE) approaches [24, 33, 50]. However, unlike these methods—which focus on improving visual quality for human perception—our Torch module aims to boost semantic representations favorable for downstream vision tasks under varying lighting conditions. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the Torch module computes an enhancement map $\mathcal{H}_\theta(\mathbf{I})$, parameterized by network weights θ. The transformed image $\mathbf{T}$ is obtained by:

$$\mathbf{T} = \mathbf{I} \odot (1 + \mathcal{H}_\theta(\mathbf{I})), \text{ where } \mathcal{H}_\theta(\mathbf{I}) = \tanh\left(\mathcal{F}(\mathbf{I};\theta)\right) \quad (1)$$

where $\odot$ denotes element-wise multiplication. The $\tanh(\cdot)$ activation ensures that the adjustment values are in the range $[-1, 1]$, allowing the model to both amplify (positive values) and attenuate (negative values) features. This formulation enables the model to adapt to the image's content effectively.

Torch network. The Torch network $\mathcal{F}(\mathbf{I};\theta)$ is a lightweight convolutional network composed of an initial feature extraction layer followed by a residual block, as illustrated in Fig. 2. The residual learning enables the Torch module to focus on learning the necessary adjustments to improve the input image rather than reconstructing it entirely. Enhancing semantically important features while preserving original content makes the Torch module crucial in maintaining and improving performance in downstream vision tasks. Notably, thanks to the formulation explained in Eq 1, our Torch module only has **11,075** trainable parameters and requires **0.712 GFLOPs** for $\mathbf{I}$ of size $256 \times 256 \times 3$, making it suitable for real-time high-level vision tasks under varying lighting conditions.

3.2. Adapt

While the Torch module enhances the input image $\mathbf{I}$ to produce $\mathbf{T}$, applying this enhancement uniformly may not be optimal. Different regions within an image require varying degrees of enhancement based on content (low/well-lit images) and context (downstream task). To address this, we introduce the **Adapt** module, which learns a content-adaptive modulation matrix in the feature space. This allows the model to selectively amplify or suppress the enhancements produced by the Torch module, improving their effectiveness for downstream vision tasks in varying illumination scenarios, effectively making it light-agnostic.

Specifically, the Adapt module computes an *Adaptive Matrix* $\mathbf{M} \in \mathbb{R}^{H\times W}$, where H and W are the spatial dimensions of $\mathbf{I}$. The purpose of the Adaptive Matrix $\mathbf{M}$ is to modulate the feature enhancements from Torch $\mathbf{T}$ before they are applied to $\mathbf{I}$. $\mathbf{M}$ is computed as:

$$\mathbf{M} = \sigma\left(\mathcal{G}_\phi(\mathbf{I})\right), \quad (2)$$

where $\mathcal{G}_\phi$ is a lightweight convolutional network parameterized by weights ϕ, and $\sigma(\cdot)$ denotes the sigmoid activation function ensuring that the values of $\mathbf{M}$ are in the range $[0, 1]$. We feed the input image $\mathbf{I}$ into the Adapt module to compute $\mathbf{M}$, ensuring that the image content directly influences the modulation.

Adapt network. As shown in Fig. 2, the convolutional network $\mathcal{G}_\phi$ in the Adapt module is designed to extract hierarchical features, which are proven to be beneficial for high-level vision tasks such as object detection and semantic segmentation [26, 45, 46]. To capture multi-scale contextual information, the network consists of three 3×3 convolutional layers with increasing channel dimensions C_1, C_2, and C_3, respectively. Each layer is followed by a batch normalization [29] and a ReLU activation function. This architecture allows the network to learn a hierarchy of features from low-level edges and textures to high-level semantic information. Finally, an additional 3×3 convolutional layer maps the features from C_3 to a single channel and applies a sigmoid activation function to generate the Adaptive Matrix $\mathbf{M} \in [0, 1]^{H\times W}$. This design balances expressiveness and computational efficiency, enabling the Adapt module to adjust enhancements based on both local and global features, which is crucial for content-adaptive modulation.

Illumination-aware scaling. We introduce a **scale factor** γ to ensure that the enhancement adapts appropriately to different illumination levels. This scale factor modulates the $\mathbf{M}$ globally based on the image's overall brightness, allowing the model to apply stronger enhancements to darker images and milder adjustments to well-lit images, maintaining visual consistency and preventing artifacts. We compute γ as a function of the average luminance $L(\mathbf{I})$ of the input image:

$$\gamma = \frac{\kappa}{1 + \exp\left(k\left(L(\mathbf{I}) - L_0\right)\right)}, \quad (3)$$

where κ is the maximum scale factor (empirically set to 3 unless stated otherwise). k controls the slope of the sigmoid function, and L_0 is the luminance value at the sigmoid's midpoint. The average luminance $L(\mathbf{I})$ is computed following the standard luminance calculation in color science [2, 42, 87].

Adaptive Fusion. The final enhanced image $\mathbf{E}$ is then obtained by modulating the enhancement from the Torch module with the illumination-aware scaled Adaptive Matrix:

$$\mathbf{E} = \mathbf{I} + ((\gamma \cdot \mathbf{M}) \odot \mathbf{T}) \,. \tag{4}$$

where $\odot$ denotes element-wise multiplication, and $\mathbf{T}$ is obtained from Eq. (1). This formulation in Eq. 4 integrates both content-based modulation (through $\mathbf{M}$) and illumination-based scaling (through γ), enabling the model to adaptively enhance images under varying lighting conditions.

3.3. Light-Agnostic Learning

Achieving robust performance across varying lighting conditions requires our model to learn representations invariant to illumination changes. Collecting labeled data under all possible lighting conditions is impractical. Hence, we adopt a self-supervised learning (SSL) strategy inspired by BYOL [23] to learn illumination-invariant features. Unlike prior works that design cumbersome pipelines to incorporate SSL into each high-level vision task [16, 48, 65], we leverage SSL to pre-train our TorchAdapt module with a pretext task that aligns backbone representations [19, 27] regardless of illumination. This approach eliminates the need for task-specific labeled data or low-light data, allowing training on general-purpose datasets like ImageNet [18]. Figure 3 illustrates the pre-training pipeline of TorchAdapt.

Objective Function. Given a well-lit image V_1, we generate its low-light counterpart $V_2 = D(V_1)$ using a darkening function $D(\cdot)$, which can be a gamma transformation, an illumination-degradation pipeline [16], or a darkening module from [48]. Let f denote the frozen pre-trained backbone network (e.g., ResNet [27], ViT [19]). We pass V_1 and V_2 through the TorchAdapt module to obtain enhanced images $\hat{V}_1$ and $\hat{V}_2$, respectively. The enhanced images are processed by the backbone f and a projector network g to obtain feature representations: $z_1' = g(f(\hat{V}_1))$ and $z_2 = g(f(\hat{V}_2))$. The target representation is obtained by passing the original well-lit image V_1 through f' and g': $y_1 = g'(f'(V_1))$. Our training objective minimizes the cosine similarity loss $\mathcal{D}_{\cos}$ between the enhanced image representations and the well-lit target representation:

$$\mathcal{L} = \left(1 - \frac{{z_1'}^\top y_1}{\|z_1'\|_2 \|y_1\|_2}\right) + \left(1 - \frac{z_2^\top y_1}{\|z_2\|_2 \|y_1\|_2}\right), \tag{5}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm of a vector. Since we employ identical, frozen backbone networks f, the loss function guides TorchAdapt to enhance the low-light image $\hat{V}_2$ so its representation aligns with that of the well-lit image V_1. Simultaneously, it discourages over-enhancement of the already well-lit image $\hat{V}_1$, as its representation should remain consistent with V_1. This training strategy enables TorchAdapt to effectively enhance low-light images while preserving well-lit ones, promoting illumination invariance in the learned representations.

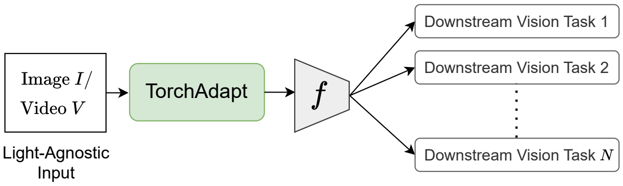


Figure 4. **TorchAdapt's downstream integration pipeline.** TorchAdapt works as a plug-and-play, backbone-agnostic feature enhancement module, seamlessly integrating into various downstream vision tasks. Given any image or video input across diverse lighting conditions, TorchAdapt produces illumination-invariant representations, significantly boosting performance without requiring task-specific modifications.

Training Details. The training of TorchAdapt is fast and efficient due to its lightweight architecture and the absence of gradient propagation through the frozen backbone network. We train TorchAdapt for 5 epochs on the ImageNet [18] dataset using a single GPU, with a learning rate of 0.0001 and a batch size of 256. Since the backbone networks are pre-trained and we aim for consistent views, we apply only color jitter and random horizontal flip as data augmentations. Pre-training TorchAdapt with a ResNet-50 backbone takes $\sim$ 6 hours on a single A100 GPU. Once trained, TorchAdapt can be seamlessly integrated into any downstream vision task, as illustrated in Fig. 4.

Dataset	Task	#Cls	#Train	#Val
Low-light (LL)				
ExDark [47]	Object detection	12	5891	1472
DARK FACE [80]	Face detection	1	5400	600
ACDC [60]	Semantic segmentation	19	400	106
LIS [10]	Instance segmentation	8	1561	669
DarkVision [81]	Video object detection	4	26*	6*
Light-Agnostic (LA)				
COCO [21] ∩ ExDark [47]	Object detection	12	92357	5156
WIDER FACE [78] ∩ DARK FACE [80]	Face detection	1	18280	3822
CityScapes [15] ∩ ACDC [60]	Semantic segmentation	19	3375	606
COCO [40] ∩ LIS [10]	Instance segmentation	8	44321	2550
ImageNet VID [58] ∩ DarkVision [81]	Video object detection	2	401*	51*

Table 1. **Datasets Statistics.** *: video samples. Light-Agnostic datasets are obtained through Eq. 6

.

4. Experiments

We conduct extensive experiments on several representative visual tasks to demonstrate the superiority of TorchAdapt in both low-light and light-agnostic settings. These tasks include object detection [40, 47, 52], face detection [78, 80], instance segmentation [10, 40], semantic segmentation [15, 60], and video object detection [58, 81]. This section first presents the experimental settings, comparing TorchAdapt

Tasks →	**Object Det.**				**Face Det.**				**Instance Seg.**				**Semantic Seg.**		**Video Object Det.**	
Light Settings →	**LL**		**LA**		**LL**		**LA**		**LL**		**LA**		**LL**	**LA**	**LL**	**LA**
Methods	mAP_{50}	mAP	mAP_{50}	mAP	AP_{50}	AP	AP_{50}	AP	mAP	SegAP	mAP	SegAP	mIoU	mIoU	mAP	mAP
Baseline	78.8	44.2	62.4	33.1	47.3	19.9	52.6	28.0	54.0	46.0	41.5	33.9	49.6	75.5	32.8	51.3
Enhancement																
Zero-DCE [24]	76.2	42.7	57.6	29.5	47.4	20.1	47.6	22.9	51.5	43.5	39.2	32.1	42.5	59.4	7.83	28.3
SCI [84]	75.5	42.2	57.9	29.7	45.1	19.3	45.4	21.7	52.3	44.2	35.1	28.3	42.1	57.4	6.71	25.7
URetinexNet [71]	74.6	42.0	57.3	29.2	44.7	18.3	46.6	22.3	59.5	43.1	37.4	29.7	41.2	58.8	6.68	25.1
IAT [17]	74.1	41.0	52.4	30.5	42.1	15.9	37.1	17.4	49.7	42.3	36.4	30.1	41.7	56.2	6.24	24.5
Task-specific																
Zero-DCE [24]‡	79.1	44.4	62.9	33.2	50.4	21.1	50.1	23.5	59.1	52.7	44.4	35.7	44.7	61.0	34.1	51.5
SCI [50]‡	79.2	44.2	63.0	33.4	51.6	22.2	54.7	29.9	58.4	51.0	43.7	34.9	43.5	59.9	32.9	50.7
MAET [16]‡	79.2	44.3	57.3	27.3	44.3	18.7	47.9	23.1	-	-	-	-	-	-	-	-
Xue *et al.* [76]	-	-	-	-	46.1	17.7	44.2	21.0	-	-	-	-	42.1	69.7	-	-
FeatEnHancer [26]‡	79.2	44.8	62.9	33.3	47.2	19.9	51.6	24.5	59.6	52.4	44.9	36.1	50.1	76.1	34.6	52.7
TorchAdapt	**80.1**	**45.4**	**64.0**	**34.1**	**53.7**	**23.1**	**59.0**	**32.5**	**61.1**	**53.6**	**45.2**	**36.6**	**50.8**	**76.3**	**39.1**	**59.1**
vs. prev. SoTA	**+0.9**	**+0.6**	**+1.0**	**+0.7**	**+2.1**	**+0.9**	**+4.3**	**+2.6**	**+1.5**	**+0.9**	**+0.3**	**+0.5**	**+0.7**	**+0.2**	**+4.5**	**+6.4**

Table 2. **Quantitative comparison across five tasks under both Low-Light (LL) and Light-Agnostic (LA) settings.** ‡: all these methods are reproduced by us with their official codebase and trained end-to-end for each task. The rest of the results are obtained from their official checkpoints. - indicates that the method is not applicable. TorchAdapt outperforms prior SOTA across all five tasks in LL and LA settings.

with baselines, existing low-light image enhancement (LLIE) methods, and state-of-the-art task-specific approaches. We then examine critical design choices for TorchAdapt through ablation studies. Finally, we validate the generalizability of TorchAdapt by comparing it with recent state-of-the-art domain adaptation methods [20, 32, 48].

Experimental Settings. We perform experiments under two settings for each task:

1)- *Low-Light (LL) Setting:* Following prior works [16, 26, 76], we utilize publicly available low-light benchmarks specific to each task. These benchmarks consist exclusively of images captured under low-light conditions.

2)- *Light-Agnostic (LA) Setting:* To introduce a more challenging and realistic evaluation, we curate datasets that encompass both well-lit and low-light images. We achieve this by selecting classes common to both well-lit and low-light datasets. Mathematically, for a given task, let $\mathcal{C}_{\text{WL}}$ and $\mathcal{C}_{\text{LL}}$ denote the sets of classes in the well-lit and low-light datasets, respectively. We define the set of common classes as $\mathcal{C}_{\text{com}} = \mathcal{C}_{\text{WL}} \cap \mathcal{C}_{\text{LL}}$. The LA dataset is then formed by combining all images from both datasets that belong to classes in $\mathcal{C}_{\text{com}}$:

$$\mathcal{D}_{\text{LA}} = \left\{ (x, y) \,\middle|\, (x, y) \in \mathcal{D}_{\text{WL}} \cup \mathcal{D}_{\text{LL}},\ y \in \mathcal{C}_{\text{com}} \right\}, \quad (6)$$

Where $\mathcal{D}_{\text{WL}}$ and $\mathcal{D}_{\text{LL}}$ are the well-lit and low-light datasets, respectively, and (x, y) represents an image and its label. We summarize the key statistics of the employed benchmarks for both settings in Table 1. *We provide complete implementation details for each experiment in Appendix A.*

Comparisons with State-of-the-Arts. We directly compare our approach with several LLIE methods, including Zero-DCE [24], IAT [17], SCI [50], and URetinexNet [71]. Following common practice [16, 26, 76], we use their released checkpoints to enhance all images before feeding them into the detector for performance evaluation. Notably, due to their zero-reference nature requiring no paired images, some of these LLIE methods [24, 50] can be trained end-to-end on downstream vision tasks. Therefore, besides evaluating their conventional enhancement results, we integrate them into each task and report their performance, similar to low-light task-specific works like MAET [16] and FeatEnHancer [26] (indicated with ‡ in Table 2). *Our findings reveal that end-to-end training of such LLIE works [24, 50]*[1] *significantly improves their performance on downstream vision tasks in both LL and LA settings.* Concurrent to our work, we notice that two new methods [28, 38] appear in the literature on improving low-light object detection. However, since their code and checkpoints were not released by the time of the submission, we could not draw direct comparisons with them. Next, we discuss experiments on each task.

4.1. Object Detection

Settings. Following common practice [16] in LL object detection, we adopt YOLOV3 [57] as the baseline detector. For the LL setting, COCO [40] pre-trained weights are utilized to fine-tune YOLOV3 on the ExDark dataset [47]. We train the model from scratch in the LA setting, as COCO images are also included in the training data (see Table. 1).

Results. As shown in Table 2, TorchAdapt achieves a superior mAP_{50} of **80.1** in LL and **64.1** in LA, compared to the baseline's 78.8 and 62.4. End-to-end training boosts the performance of enhancement methods like Zero-DCE [33] and SCI [50]; however, TorchAdapt still surpasses them, and the low-light object detection method FeatEnHancer [26] by **+0.9** and **+1.1** in LL and LA settings, respectively. These results affirm the efficacy of TorchAdapt for object detec-

[1] We select these works due to their efficiency and reproducibility. Experiments with more methods are in Appendix C.

tion across varying illumination conditions. We provide a qualitative comparison in Appendix B.

4.2. Face Detection

Settings. For face detection, we choose a different detector and adopt RetinaNet [41] to report results on the DARK FACE dataset [69, 80] in the LL setting and a combination of the DARK FACE and WIDER FACE [78] datasets in the LA setting. Following prior works [26], we resize images to a resolution of 1500 × 1000 pixels and follow the 1× schedule in MMDetection [9].

Results. Results in Table 2 reveal that TorchAdapt demonstrates superior performance on face detection under both LL and LA settings, outperforming the previous best method (SCI‡ [50]) by **+2.1** and **+4.3** in mAP_50, respectively. The significant gains, especially in the challenging LA setting, highlight TorchAdapt's capacity to handle diverse illumination scenarios by adaptively modulating features based on content and illumination, thereby improving the robustness and accuracy of the baseline. We also compare TorchAdapt with recent domain adaptation methods [20, 48] using DSFD [36] in Sec. 4.7.

4.3. Instance Segmentation

Settings. For the instance segmentation task, we adopt RTMDet-Ins-tiny [49][2] as our baseline model and resize images to 640 × 640 pixels for training. Following Section 4.1, we use the COCO pre-trained weights of RTMDet-Ins-tiny for the LL setting and train the model from scratch for the LA setting (see Table 1).

Results. As summarized in Table 2), TorchAdapt achieves SOTA SegAP performance in both low-light (LL) and light-agnostic (LA) settings. Under LL conditions, TorchAdapt attains a SegAP of **53.6**, surpassing the previous best (FeatEnHancer‡) by **+1.2**. In the LA setting, it achieves a SegAP of **36.6**, slightly improving over FeatEnHancer‡ by **+0.5**. These results demonstrate TorchAdapt's effectiveness in enhancing instance segmentation across varying illumination by producing illumination-invariant features.

4.4. Semantic Segmentation

Settings. Consistent with prior works [26, 76], we adopt DeepLabV3+[11] as our baseline model and follow [26, 76] for the LL setting. For the LA evaluation (Table 1), we train DeepLabV3+ from scratch using MMSegmentation [14].

Results. As shown in Table 2, TorchAdapt achieves an mIoU of **50.8** in the LL setting, surpassing the previous best (FeatEnHancer‡) by **+0.7**, and an mIoU of **76.3** in the LA setting, improving over FeatEnHancer‡ by **+0.2**. These results highlight TorchAdapt's effectiveness for semantic segmentation under varying illumination conditions.

[2] https://github.com/open-mmlab/mmdetection/blob/main/configs/rtmdet/

#	Torch(§ 3.1)	Adapt(§ 3.2)	Scaling(§ 3.2)	Object Det.		Instance Seg.	
				LL	LA	LL	LA
1	×	×	×	78.8	62.4	46.0	33.9
2	✓	×	×	79.1	62.9	51.1	34.8
3	✓	✓	×	79.9	63.5	53.3	36.2
4	✓	×	✓	79.5	62.7	51.4	34.1
5	✓	✓	✓	**80.1**	**64.0**	**53.6**	**36.6**

Table 3. **Contribution of each component.** Highlighted settings are set as default.

Scale factor (κ)	OD		SS	
	LL	LA	LL	LA
2	79.9	63.7	53.4	36.4
3	80.1	64.0	53.6	36.6
4	79.9	63.3	53.2	36.2
5	79.5	61.1	53.3	34.5

(a) **Scale factor (κ) in Eq. 3.**

$D(\mathbf{I})$	OD		SS	
	LL	LA	LL	LA
ISP [16]	79.4	61.7	47.6	29.9
$D(\mathbf{I})$ in [48]	79.6	63.1	53.2	36.3
Gamma curve	80.1	64.0	53.6	36.6

(b) **$D(\mathbf{I})$ in § 3.3 and Fig. 3.**

Table 4. **Ablating design choices in TorchAdapt.** Highlighted settings are set as default. OD and SS are object detection and semantic segmentation, respectively.

4.5. Video Object Detection

Settings. In addition to image-based tasks, we evaluate the generalizability of TorchAdapt on video data. We present the dataset details in Table 1. Other experimental settings follow [26], with additional implementation details provided in Appendix A.

Results. We report the results of the challenging video object detection task in Table 2. While all LLIE methods perform poorly on this task, end-to-end training significantly boosts the performance of zero-reference methods [24, 50]. TorchAdapt, however, outperforms all methods under both LL and LA settings, achieving substantial gains of **+4.5** and **+6.4**, respectively. These improvements affirm TorchAdapt's robustness and effectiveness for video object detection across both low-light and light-agnostic settings.

4.6. Ablation Studies

We ablate the components and design choices of TorchAdapt in object detection (OD) and instance segmentation (IS) tasks under (Low-light) LL and (Light-Agnostic) LA settings, following the same implementation details in § 4.1 and § 4.3, respectively. We provide more ablations in Appendix C.

Contribution of Each Component. Table 3 evaluates the impact of the Torch module, the Adapt module, and Illumination Scaling. Adding the Torch module alone (row 2) improves performance over the baseline with **+0.3%** in LL OD and **+5.1%** in LL IS, highlighting the benefit of feature enhancement in low-light vision. Introducing the Adapt module alongside Torch (row 3) yields significant improvements (**+0.8%**/**+0.6%** in LL/LA OD; **+2.2%**/**+1.4%** in LL/LA IS), underscoring the importance of content-adaptive modulation. Employing scaling without the Adapt module (Row 4) offers minor gains in LL settings. However, it is detrimental in LA settings (**-0.2%**/**-0.7%** in OD/IS), indicating that content-based adaptation is more impactful than global scaling alone. The full TorchAdapt configuration, combining all three components (row 5), achieves the best results with

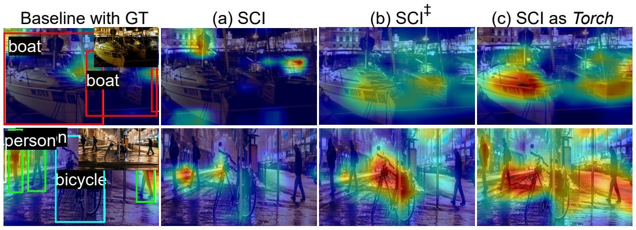


Figure 5. **Visualizing backbone features with SCI [50]** on LA object detection using YOLOV3. **(a)** SCI's official checkpoint as an enhancer. **(b)** SCI† trained end-to-end. **(c)** SCI integrated into our TorchAdapt framework as a Torch. SCI with TorchAdapt extracts more representative features in both low and well-lit images.

Method	**mAP (%)**
DSFD [36]	16.1
CICony [32]	18.4
SimMinMax [48]	25.7
DAI-Net [20] ‡	28.0
DAI-Net+TorchAdapt	**31.9+3.9**

Table 5. **Zero-shot performance comparison of TorchAdapt with state-of-the-art domain adaptation methods** on DARK FACE [80] using DSFD [36]. Results with ‡ are reproduced using the original codebase. TorchAdapt enhances prior SOTA [20] by +3.9 mAP, demonstrating its effectiveness in zero-shot domain adaptation.

+1.3%/+1.6% in LL/LA OD and **+7.6%/+2.5%** in LL/LA IS over the baseline, confirming that these components complement each other and synergistically enhance performance in both LL and LA settings.

Design Choices in TorchAdapt. Table 4 ablates important design choices in TorchAdapt. As shown in Table 4a, we empirically establish that when the scale factor γ in Eq. 3 is computed with $\kappa = 3$, it produces optimal performance. Results in Table 4b reveal that employing the Gamma Curve in $D(I)$ to generate V_2 in § 3.3 brings the biggest boost in performance across both LL and LA settings.

4.7. Comparison with Domain-Adaptation Methods

We evaluate TorchAdapt's effectiveness for zero-shot domain adaptative face detection on the DARK FACE dataset [80]. Following settings from [20], we train the detector only on the WIDER FACE dataset [78]. As shown in Table 5, TorchAdapt integrated with DAI-Net [20], the current best-performing method, achieves a significant improvement of +3.9% mAP, demonstrating its effectiveness in enabling robust zero-shot adaptation under diverse illumination conditions.

5. Exploring TorchAdapt's Properties

Model-Agnostic Generality. Experiments in § 4 confirm that TorchAdapt is a general-purpose module that can be integrated into any high-level vision task. Moreover, TorchAdapt as a framework is highly flexible: any zero-reference (unsupervised) LLIE methods [33, 50, 85] can serve as the *Torch* component to improve performance under varying illuminations. To verify this, we adopt SCI [50] as the Torch module in TorchAdapt and visualize the learned backbone features in Fig.5(c). Using SCI within TorchAdapt yields highly representative backbone features for both low-light and well-lit images. Appendix C provides additional examples and quantitative analyses.

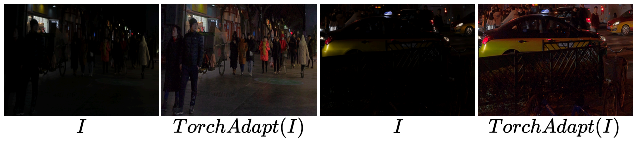


Figure 6. Illustrating **low-light image Enhancement as an emergent property of TorchAdapt** on the DARK FACE validation set. More results can be found in Appendix B.

Method	mAP$_{50}$(LL)	mAP$_{50}$(LA)	#Params.(M)	Latency (ms)
Baseline [57]	78.8	62.4	65.2	21.9
+TorchAdapt	**80.1** (+1.3)	**64.0** (+1.6)	**65.3** (+0.1)	23.7 (+1.8)

Table 6. **Computational analysis**. With YOLOV3 [57], TorchAdapt introduces minimal computational overhead while bringing stronger gains across both LL and LA settings.

Enhancement as an Emergent Property. Although TorchAdapt is designed to produce illumination-invariant features for high-level vision tasks without employing any explicit enhancement loss functions during training, we observe that it inherently enhances low-light images. We illustrate this effect in Fig. 6. This emergent behavior likely results from the pre-training objective of aligning representations between well-lit and low-light images (see § 3.3), which encourages the network to implicitly adjust low-light images toward well-lit representations.

Computational Analysis. We assess the computational efficiency of TorchAdapt using YOLOV3 [57] on the object detection task. As shown in Table 6, integrating TorchAdapt incurs minimal additional computational cost while delivering significant performance improvements. This demonstrates that TorchAdapt is compatible with lightweight models and suitable for real-time applications.

6. Conclusions

In this work, we introduce TorchAdapt, a real-time, general-purpose, light-agnostic adaptive feature enhancement framework that boosts the performance of high-level vision tasks under varying illumination conditions. By integrating its modules with light-agnostic learning, TorchAdapt produces powerful semantic representations for both low-light and well-lit data. Extensive experiments on five representative visual perception tasks involving images and videos confirm its effectiveness.

7. Acknowledgement

This work was partially funded by the European Union under Horizon Europe, grant number 101135724 (LUMINOUS).

References

[1] Muhamamd Zeshan Afzal, SK Aziz Ali, Didier Stricker, Peter Eisert, Anna Hilsmann, Daniel Perez-Marcos, Marco Bianchi, Sonia Crottaz-Herbette, Roberto De Ioris, Eleni Mangina, et al. Next generation xr systems-large language models meet augmented and virtual reality. *IEEE computer graphics and applications*, 2025. 1

[2] Tadahiro Azetsu and Noriaki Suetake. Hue-preserving image enhancement in cielab color space considering color gamut. *Optical Review*, 26:283–294, 2019. 5

[3] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. Revisiting active perception. *Autonomous Robots*, 42:177–196, 2018. 1, 2

[4] Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC, 2015. 1

[5] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[7] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pages 2722–2730, 2015. 1, 2

[8] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 2

[9] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7

[10] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 2, 3, 5

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 7

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2

[13] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1

[14] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020. 7

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2, 5

[16] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2553–2562, 2021. 1, 2, 3, 5, 6, 7

[17] Ziteng Cui, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, Zhengkai Jiang, Yu Qiao, and Tatsuya Harada. You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. *arXiv preprint arXiv:2205.14871*, 2022. 1, 6

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 3, 5

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 5

[20] Zhipeng Du, Miaojing Shi, and Jiankang Deng. Boosting object detection with zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12666–12676, 2024. 1, 2, 3, 6, 7, 8

[21] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 2, 5

[22] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 3, 5

[24] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. *CoRR*, abs/2001.06826, 2020. 1, 2, 4, 6, 7

[25] Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2rnet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation*, 90:103712, 2023. 2

[26] Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammad Zeshan Afzal. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6725–6735, 2023. 1, 2, 3, 4, 6, 7

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 2, 5

[28] Mingbo Hong, Shen Cheng, Haibin Huang, Haoqiang Fan, and Shuaicheng Liu. You only look around: Learning illumination invariant feature for low-light object detection. *arXiv preprint arXiv:2410.18398*, 2024. 6

[29] Sergey Ioffe. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[30] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30: 2340–2349, 2021. 2

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[32] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C. van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4399–4409, 2021. 1, 2, 3, 6, 8

[33] Chongyi Li, Chunle Guo Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 4, 6, 8

[34] Chongyi Li, Chunle Guo, Shangchen Zhou, Qiming Ai, Ruicheng Feng, and Chen Change Loy. Flexicurve: Flexible piecewise curves estimation for photo retouching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1092–1101, 2023. 2

[35] Hebei Li, Jin Wang, Jiahui Yuan, Yue Li, Wenming Weng, Yansong Peng, Yueyi Zhang, Zhiwei Xiong, and Xiaoyan Sun. Event-assisted low-light video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3250–3259, 2024. 3

[36] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5060–5069, 2019. 7, 8

[37] Jinlong Li, Baolu Li, Zhengzhong Tu, Xinyu Liu, Qing Guo, Felix Juefei-Xu, Runsheng Xu, and Hongkai Yu. Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15205–15215, 2024. 2

[38] Yunan Li, Yihao Zhang, Shoude Li, Long Tian, Dou Quan, Chaoneng Li, and Qiguang Miao. Watching it in dark: A target-aware representation learning framework for high-level vision tasks in low illumination. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 6

[39] Jinxiu Liang, Jingwen Wang, Yuhui Quan, Tianyi Chen, Jiaying Liu, Haibin Ling, and Yong Xu. Recurrent exposure generation for low-light face detection. *IEEE Transactions on Multimedia*, 24:1609–1621, 2022. 3

[40] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 1, 2, 3, 5, 6

[41] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 7

[42] Hewei Liu, Renwei Yang, Shuyuan Zhu, Xing Wen, and Bing Zeng. Luminance-guided chrominance image enhancement for hevc intra coding. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 3180–3184. IEEE, 2022. 5

[43] Jiaying Liu, Dejia Xu, Wenhan Yang, Minhao Fan, and Haofeng Huang. Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision*, 129:1153–1184, 2021. 2

[44] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. 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*, pages 10561–10570, 2021. 1, 2

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 4

[46] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022. 4

[47] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *CoRR*, abs/1805.11227, 2018. 1, 2, 3, 5, 6

[48] Rundong Luo, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Similarity min-max: Zero-shot day-night domain adaptation.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8104–8114, 2023. 1, 2, 3, 5, 6, 7, 8

[49] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 7

[50] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5637–5646, 2022. 1, 2, 4, 6, 7, 8

[51] Tengyu Ma, Long Ma, Xin Fan, Zhongxuan Luo, and Risheng Liu. PIA: parallel architecture with illumination allocator for joint enhancement and detection in low-light. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 2070–2078. ACM, 2022. 1, 3

[52] Igor Morawski, Yu-An Chen, Yu-Sheng Lin, and Winston H. Hsu. NOD: taking a closer look at detection under extreme low-light conditions with night object detection dataset. *CoRR*, abs/2110.10364, 2021. 3, 5

[53] Lukáš Neumann, Michelle Karg, Shanshan Zhang, Christian Scharfenberger, Eric Piegert, Sarah Mistr, Olga Prokofyeva, Robert Thiel, Andrea Vedaldi, Andrew Zisserman, et al. Nightowls: A pedestrians at night dataset. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part I 14*, pages 691–705. Springer, 2019. 3

[54] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 3

[55] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[56] Qingpao Qin, Kan Chang, Mengyuan Huang, and Guiqing Li. Denet: Detection-driven enhancement network for object detection under adverse weather conditions. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 2813–2829, 2022. 3

[57] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 6, 8

[58] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1, 2, 5

[59] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7374–7383, 2019. 1, 2

[60] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10765–10775, 2021. 1, 2, 3, 5

[61] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark - domain adaptation method for merging multiple models. In *Computer Vision – ECCV 2020*, pages 345–359, Cham, 2020. Springer International Publishing. 3

[62] Shangquan Sun, Wenqi Ren, Tao Wang, and Xiaochun Cao. Rethinking image restoration for object detection. *Advances in Neural Information Processing Systems*, 35:4461–4474, 2022. 3

[63] Maria Valera and Sergio A Velastin. Intelligent distributed surveillance systems: a review. *IEE Proceedings-Vision, Image and Signal Processing*, 152(2):192–204, 2005. 1, 2

[64] Wenjing Wang, Wenhan Yang, and Jiaying Liu. Hla-face: Joint high-low adaptation for low light face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16195–16204, 2021. 3

[65] Wenjing Wang, Zhengbo Xu, Haofeng Huang, and Jiaying Liu. Self-aligned concave curve: Illumination enhancement for unsupervised adaptation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 2617–2626, 2022. 2, 3, 5

[66] Wenjing Wang, Xinhao Wang, Wenhan Yang, and Jiaying Liu. Unsupervised face detection in the dark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1250–1266, 2023. 3

[67] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26057–26066, 2024. 2

[68] Xinzhe Wang, Kang Ma, Qiankun Liu, Yunhao Zou, and Ying Fu. Multi-object tracking in the dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 382–392, 2024. 3

[69] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *CoRR*, abs/1808.04560, 2018. 7

[70] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 2

[71] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. 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 (CVPR)*, pages 5901–5910, 2022. 1, 6

[72] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021. 1, 2

[73] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17714–17724, 2022. 2

[74] Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9893–9903, 2023. 2

[75] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *Deep Learning for Human Activity Recognition: Second International Workshop, DL-HAR 2020, Held in Conjunction with IJCAI-PRICAI 2020, Kyoto, Japan, January 8, 2021, Proceedings 2*, pages 70–84. Springer, 2021. 3

[76] Xinwei Xue, Jia He, Long Ma, Yi Wang, Xin Fan, and Risheng Liu. Best of both worlds: See and understand clearly in the dark. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 2154–2162, New York, NY, USA, 2022. Association for Computing Machinery. 6, 7

[77] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1

[78] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 1, 2, 3, 5, 7, 8

[79] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12918–12927, 2023. 2

[80] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, Yuqiang Zheng, Yanyun Qu, Yuhong Xie, Liang Chen, Zhonghao Li, Chen Hong, Hao Jiang, Siyuan Yang, Yan Liu, Xiaochao Qu, Pengfei Wan, Shuai Zheng, Minhui Zhong, Taiyi Su, Lingzhi He, Yandong Guo, Yao Zhao, Zhenfeng Zhu, Jinxiu Liang, Jingwen Wang, Tianyi Chen, Yuhui Quan, Yong Xu, Bo Liu, Xin Liu, Qi Sun, Tingyu Lin, Xiaochuan Li, Feng Lu, Lin Gu, Shengdi Zhou, Cong Cao, Shifeng Zhang, Cheng Chi, Chubing Zhuang, Zhen Lei, Stan Z. Li, Shizheng Wang, Ruizhe Liu, Dong Yi, Zheming Zuo, Jianning Chi, Huan Wang, Kai Wang, Yixiu Liu, Xingyu Gao, Zhenyu Chen, Chang Guo, Yongzhou Li, Huicai Zhong, Jing Huang, Heng Guo, Jianfei Yang, Wenjuan Liao, Jiangang Yang, Liguo Zhou, Mingyue Feng, and Likun Qin. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020. 1, 2, 3, 5, 7, 8

[81] Bo Zhang, Yuchen Guo, Runzhao Yang, Zhihong Zhang, Jiayi Xie, Jinli Suo, and Qionghai Dai. Darkvision: A benchmark for low-light image/video perception, 2023. 1, 2, 3, 5

[82] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[83] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2547–2555, 2019. 3

[84] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. *CoRR*, abs/1905.04161, 2019. 1, 2, 6

[85] Zhaoyang Zhang, Yitong Jiang, Jun Jiang, Xiaogang Wang, Ping Luo, and Jinwei Gu. Star: A structure-aware lightweight transformer for real-time image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4106–4115, 2021. 1, 8

[86] Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, and Meng Wang. Deep color consistent network for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1899–1908, 2022. 2

[87] Hegui Zhu, Kai Wang, Ziwei Zhang, Yuelin Liu, and Wuming Jiang. Low-light image enhancement network with decomposition and adaptive information fusion. *Neural Computing and Applications*, 34(10):7733–7748, 2022. 5