BRNet: a Bio-Receptor Network for Object Detection with Zero-Shot Domain

Chanhee Lee Byeongho Ko* Jingwoong Jung* Yeonhoo Jung * Jaekwang Kim†
Department of Applied Artificial Intelligence
Sungkyunkwan University
{leechanhye, polite337, dusgnwjd, wjdwlsdndpp}@(g.)skku.edu

Abstract

Visual recognition in low-light environments remains a challenging task, as detectors trained on well-lit data often fail under poor visibility, low contrast, and severe illumination shifts. To overcome this, we present BRNet, a biologically inspired detector that adaptively modulates feature extraction in response to ambient brightness. The core component of BRNet is the Photo Receptor module, which emulates retinal rod and cone cells to adaptively extract contrast-sensitive or semantic features depending on brightness levels. We derive a luminance-based dark-level estimation function grounded in mesopic vision theory to pseudo-label brightness levels, guiding the dynamic activation of the Rod and Cone pathways. To prevent interference between detection and auxiliary tasks such as reflectance and darkness prediction, we introduce a Semi Orthogonal Loss that selectively decorrelates overlapping feature subspaces while preserving shared semantics. BRNet achieves strong generalization on DARK FACE and ExDark under a zero-shot day-to-night adaptation setting, without requiring image enhancement or retraining. The code is available at https://github.com/iontail/BRNet.git.

1. Introduction

Object detection in low-light conditions remains a fundamental challenge in computer vision. Models trained on standard datasets often fail to generalize under poor illumination, where low contrast, noise, and extreme lighting variations obscure object boundaries and degrade feature representations. These failures are not solely due to data scarcity but also because existing detectors are not inherently designed to adapt across diverse illumination scenarios.

Traditional approaches have attempted to resolve this issue through image enhancement techniques [8, 38], leveraging Retinex theory or curve estimation to improve visual quality. Others employ unsupervised domain adaptation (UDA) [18, 40], aiming to bridge the gap between day and night domains by translating appearance or aligning features. However, both strategies are limited: enhancement methods often neglect task-specific objectives like detection accuracy, while domain adaptation methods struggle with generalization across real-world lighting variability.

To move beyond these limitations, we draw inspiration from the biological visual system. In particular, the retina dynamically balances rod and cone cell activations to handle extreme luminance variation. Motivated by this mechanism, we propose **BRNet (Bio-Receptor Network)**, a novel object detector designed to adaptively extract illumination-aware features. At its core is the *Photo Receptor module*, composed of Rod and Cone blocks. The Rod block focuses on extracting contrast-sensitive features under darkness using a Gain module, a Tapetum-inspired structure, and deformable convolution. The Cone block operates in brighter conditions to capture spatial semantics. A biologically motivated *dark-level function*, based on mesopic luminance theory, modulates the balance between these two pathways depending on the brightness of the input.

To further improve robustness and mitigate task interference, we introduce a **Semi Orthogonal Loss**. This feature-level regularization partial disentangles shared representations between the main detection task and auxiliary branches (e.g., reflectance or brightness estimation). By avoiding full orthogonal constraints, it maintains useful semantic overlap while reducing harmful entanglement.

Our proposed BRNet operates effectively without explicit image enhancement or additional nighttime supervision. It achieves strong generalization from day to night and across variable lighting conditions, demonstrating that biologically inspired architectures, when combined with carefully designed multi-task regularization, can significantly improve low-light object detection. Our contributions are summarized as follows:

- We propose BRNet, a biologically inspired object detector that adaptively extracts contrast or semantic features through photoreceptor-mimicking pathways.

[*] Equal contribution (co-second authors).
[†] Corresponding author

- We design a dark-level modulation function derived from mesopic luminance theory to dynamically regulate Rod and Cone activations.
- We introduce a feature-level Semi Orthogonal Loss to reduce task interference while preserving beneficial feature sharing in a multi-branch architecture.

2. Related Work

2.1. Object Detection in the Low-Light Condition

Object detection under low-light condition has been a challenging problem in many fields, due to its sparse information compared with well-lit condition. Thanks to the improvement of deep learning, many methods tried to deal with this low-light problems. According to the Retinex theory [17], when the human visual system perceives a scene, it does not recognize the absolute brightness at a specific location, but rather the relative brightness in comparison to the surrounding areas. Reflectance and illumination is the components of an image, and many other works [15, 22, 26, 36, 38, 39, 43] also applied techniques to make improvement, using the characteristic of this theory. However, these researches couldn't benefit the understanding of the images.

One another approach to improving low-light object detection performance was to enhance the images fed into the detection algorithm. Wei et al. [38] adopted deep learning image enhancement based on Retinex theory. EnlightenGAN [14] introduced GAN architecture to deal with various unpaired real images, and Zero-DCE [8] focused on curve estimation based unsupervised learning. In addition, many other image enhancement studies have been conducted [1, 27, 32, 46]. More recently, various approaches have been explored to improve distinctive features relevant to object detection, rather than just enhancing the image itself [10, 12, 28, 41, 45]. PE-YOLO [45] combines PENet with YOLOv3 [31], which can extract the feature of various resolution effectively. FeatEnhancer [10] proposed a method that enhances the representation of low-light images by hierarchically integrating multi-scale features and learning through task-related loss functions.

Meanwhile, there have been proposed several dataset with low-light object detection task. With other normal condition image datasets [13, 21, 42], Neumann et al. [29] suggested Nightowls, which includes various pedestrian images in night environment. Loh et al. [24] introduced the ExDark dataset, consisting of 7,363 images captured under varying low-light conditions ranging from very dark to twilight, annotated with 12 object categories. While such datasets are well-suited for evaluating object detection under challenging illumination, they lack coverage of well-lit scenarios, limiting their applicability in models targeting broader lighting conditions.

2.2. Domain Adaptation

An alternative direction to address the data scarcity problem in low-light detection is unsupervised domain adaptation(UDA) from day to night. Instead of collecting large annotated night datasets, UDA methods train on labeled daytime images and adapt the detector to unlabeled nighttime images. Early attempts used image-level style transfer to generate synthetic nighttime training data from daytime photos, enabling the detector to see "night-like" imagery during training. More advanced techniques employed feature-level adaptation and self-training. For example, Cycle GAN variants and style transfer were used to preserve object identity during translation [40], and teacher-student frameworks with consistency losses were applied to gradually adapt detectors to dark conditions.

Extending beyond UDA, zero-shot day-to-night adaptation has also been explored, where a model trained solely on day images is adapted without seeing any real dark images at all. Lengyel et al. [18] introduced a physics prior-based zero-shot adaptation.

These domain adaptation strategies have significantly reduced the reliance on curated low-light datasets by leveraging abundant daytime data and unlabeled nighttime images, or even synthetic images generated through physical simulation. Despite these advances, a noticeable performance gap remains between adaptation-based approaches and fully supervised models, highlighting the need for continued research to more effectively bridge the day-to-night domain gap.

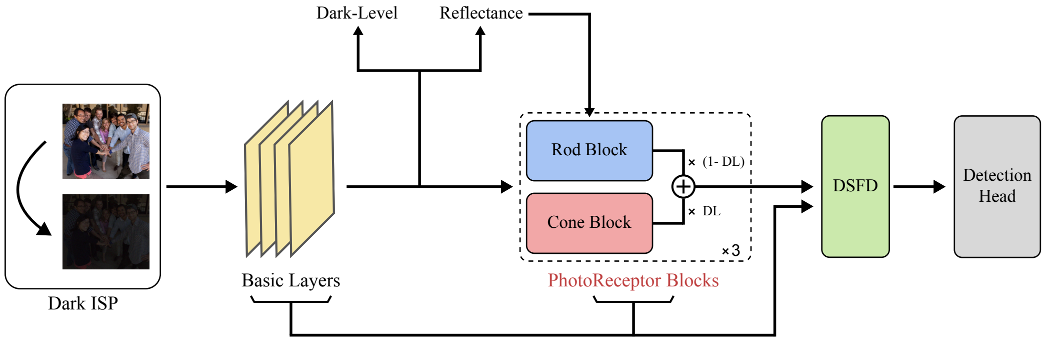
2.3. Multi-Task Learning and Representation Entanglement

Multi-task learning can enhance model performance by incorporating auxiliary prediction branches. However, it also poses the risk of representation entanglement, where shared layers become dominated by auxiliary task semantics, ultimately hindering the optimization of the primary task.

To address this issue, DMTRL [44] models inter-task relationships through tensor factorization, enabling soft parameter sharing in a data-driven manner. MTAN [23] introduces task-specific attention masks over a shared backbone, allowing each task to focus on relevant features and reduce interference. MAET [3] proposes an orthogonal loss to decorrelate task-specific gradients, motivated by the high entanglement risk when predicting multiple degradation parameters simultaneously. In contrast, DAI-Net [6] does not apply disentanglement strategies, since it only includes reflectance prediction in the auxiliary branch. As a result, the risk of feature interference is relatively low.

Building upon these insights, we design a tailored regularization strategy for our architecture, which includes an additional dark-level prediction branch beyond the re-

Figure 1. Overall architecture of BRNet. The network predicts reflectance and dark-level values through separate branches, supervised by labels generated from the Dark ISP module [3]. These predictions guide the Photo Receptor blocks to extract features that are robust to domain shifts between low-light and well-lit conditions. Final detection is performed by a DSFD-based head.



flectance estimation employed in DAI-Net. This added complexity raises the risk of task interference and representational entanglement, making naïve parameter sharing suboptimal. To address this, we introduce a Semi Orthogonal Loss that selectively enforces orthogonality between task-specific feature subspaces, rather than across the entire representation space.

This partial disentanglement ensures that task-specific representations do not interfere destructively, while shared components are still co-learned effectively. Combined with biologically inspired modulation in our Photo Receptor blocks, this formulation enables the model to retain semantic alignment across tasks while enhancing robustness under diverse lighting conditions.

3. Method

3.1. Overall Architecture

We propose a domain-adaptive object detection framework tailored for both low-light and well-lit environments. As illustrated in Figure 1, the model consists of three main components: a Dark ISP module, a set of Photo Receptor blocks, and a DSFD-based detection head. The Dark ISP module generates pseudo-labels for reflectance, which are used to supervise one auxiliary branch. In addition, we introduce a dark-level estimation function that provides pseudo-labels for illumination levels based on mesopic luminance theory. These dark-level values supervise the second auxiliary branch and are used to modulate the activation of the Photo Receptor blocks. The Photo Receptor blocks, composed of Rod and Cone sub-blocks, adaptively extract either high-frequency contrast or low-frequency semantic features depending on the predicted dark-level. This biologically inspired modulation allows the model to extract robust representations across diverse lighting conditions. Finally, the DSFD-based detection head performs object localization and classification using the adaptively processed feature maps.

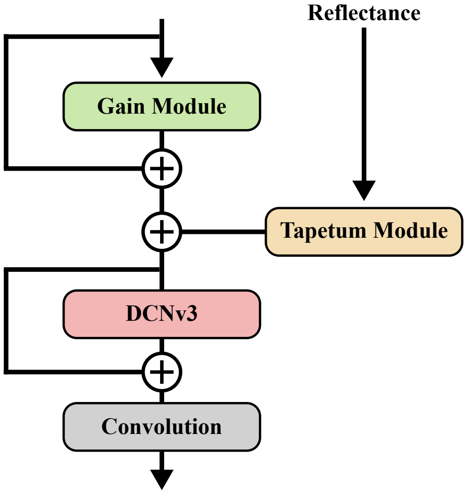
3.2. Photoreceptor Block

To achieve the main objective of dynamically regulating the activation levels of photoreceptor cells, the predicted dark-level value is used within the Photo Receptor blocks. Each Photo Receptor block is composed of two distinct components: the Rod block and the Cone block. The Rod block is designed to detect fine details and contrast in environments with low illumination, while the Cone block focuses on extracting broader spatial features under bright lighting conditions.

The Cone block consists of several standard convolution layers. In comparison, the Rod block employs a more elaborate architecture because extracting meaningful information from dark regions is significantly more challenging. Drawing inspiration from the biological visual systems of animals—such as owls, cats, and humans, which have approximately 300 times, 120 times, and 15 times more rod cells than cone cells [4], respectively -we allocate a greater number of parameters to the Rod block. This block includes three key components: the Gain module, the Tapetum module, and a deformable convolutional layer referred to as DCNv3 [37] as shown in Figure 2.

The Gain module intensifies input signals by adding them to the residual connection [11]. This design amplifies specific feature channels, enabling subsequent blocks to better capture high-frequency information. In parallel, the Tapetum module mimics the tapetum lucidum, a biological structure that reflects light behind the retina, allowing rod cells to detect more photons. The Tapetum module receives the predicted reflectance values from the auxiliary branch and further amplifies the input signals, as reflectance information is critical for object detection in low-light environments based on Retinex Theory.

Figure 2. Structure of the Rod block in the Photo Receptor block. It consists of a Gain block, Tapetum block, and DCNv3. The plus symbol (+) denotes element-wise summation.



Following these modules, the deformable convolution affiliation layer [5, 35, 37] aims to extract high frequency details. While conventional convolutional neural networks (CNNs) rely on fixed kernel sampling grids, which constrain their receptive fields to rigid and grid aligned patterns, deformable convolutions introduce learnable offsets that enable adaptive spatial sampling. This flexibility allows the network to attend to structures that are not aligned to the grid and are spatially discontinuous [5], which is particularly important under poor lighting conditions and when images suffer from degradation. As a result, deformable convolutions enhance the model's ability to capture fine grained information by relaxing locality constraints and enabling sampling behavior that is invariant to geometric transformations. Furthermore, motivated by the finding that the residual learning mechanism facilitates the model's ability to capture high frequency components [20], a residual connection is incorporated into the DCNv3 structure.

To disentangle the representations between the photoreceptor block and the dark-level prediction branch, we apply PyTorch's detach operation to block gradient flow from the dark-level outputs to the layers following the branching point in the main network. This prevents the dark-level prediction from interfering with the feature learning process in the photoreceptor block, thereby preserving task-specific representations.

3.3. Semi Orthogonal Loss

BRNet is designed to predict both reflectance and dark-level values through auxiliary branches that complement the main object detection task. While this multi-branch structure enhances the model's ability to capture illumination-aware representations, it also increases the risk of representation entanglement, where overlapping semantic representations across tasks may interfere with the optimization of the primary task. Applying full orthogonal constraints between task-specific features may mitigate this interference, but such strict disentanglement can be overly restrictive, especially when auxiliary tasks share semantically aligned cues with the main task.

To mitigate the adverse effects of representation entanglement while preserving the benefits of multi-task learning, we introduce Semi Orthogonal Loss, a selective regularization strategy that balances disentanglement and task synergy. Specifically, this loss constrains only a subset of feature channels from each auxiliary task branch to be orthogonal to those of the main detection task. Unlike a full orthogonality constraint, which may eliminate useful shared information, our approach selectively suppresses interfering signals while retaining semantically aligned components across tasks.

Originally, the semi orthogonal loss was implemented by comparing the gradients of encoder representations across tasks as in [3]:

$$\mathcal{L}_{\text{s-ort}} = \sum_{k\in\mathcal{K}, i\in\mathcal{P}} \left| \frac{\left(\frac{\partial E}{\partial T^k_{\text{aux}_i}}\right)^T \cdot \left(\frac{\partial E}{\partial T^k_{\text{obj}}}\right)}{\left|\frac{\partial E}{\partial T^k_{\text{aux}_i}}\right| \cdot \left|\frac{\partial E}{\partial T^k_{\text{obj}}}\right|} \right| \tag{1}$$

$$\text{where} \quad \mathcal{K} \subset \{1, 2, \ldots, C\}, \quad |\mathcal{K}| = \lambda C$$

However, this gradient-level approach has three key limitations. First, gradient-based comparisons can be unstable in deep multi-branch networks, especially when internal backpropagation paths differ in complexity or dynamics. Second, gradients are not available during inference, which prevents us from verifying the persistence of disentanglement. Lastly, gradient hooking imposes additional memory and computational overhead during training.

To address these issues, we adopt a feature-level formulation of the Semi Orthogonal Loss, where latent representations are directly compared via cosine similarity. This alternative offers several advantages over the gradient-based formulation. First, it provides more direct control over the dissimilarity between task-specific representations. While gradient-level orthogonality merely regularizes the learning signal, feature-level loss explicitly encourages the learned features themselves to occupy distinct subspaces, thereby improving semantic disentanglement. Second, computing cosine similarity between features during the forward pass improves training stability, as it avoids the noise and variance often introduced by gradients. Finally, this approach aligns with recent findings in the literature: methods such as BiaSwap [16], BendVLM [7], and DeCLIP [33] have shown that enforcing decorrelation or orthogonality at the feature level is an effective and practical way to promote disentangled representations in multi-task and self-supervised learning settings.

Given that our encoder features typically have smaller dimensionality than decoder outputs (e.g., DSFD head), we constrain only the *minimum* dimensional subset between tasks. Formally, let $\mathbf{f}_{\text{aux}}, \mathbf{f}_{\text{obj}} \in \mathbb{R}^{B\times D}$ be the auxiliary and main task features respectively, and let $D' = \min(D_{\text{aux}}, D_{\text{obj}})$ be the comparison between subset of dimensions. The feature-level semi orthogonal loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{s-ort}} = \lambda_{\text{sort}} \cdot \cos\left(\mathbf{f}_{\text{aux}}^{(D')},\ \mathbf{f}_{\text{obj}}^{(D')}\right) \\ + \lambda_{\text{sort-m}} \cdot \left(\left[1-\cos\left(\mathbf{f}_{\text{aux}}^{(D')},\ \mathbf{f}_{\text{aux}}^{(D')}\right)\right]\right. \\ \left. + \left[1-\cos\left(\mathbf{f}_{\text{obj}}^{(D')},\ \mathbf{f}_{\text{obj}}^{(D')}\right)\right]\right)\end{aligned} \tag{2}$$

Here, the first term penalizes alignment between the auxiliary and main task features to encourage orthogonality across tasks. The second term promotes self-consistency within each feature set by maximizing their intra-feature cosine similarity. λ_{sort} and $\lambda_{\text{sort-m}}$ are weighting coefficients that balance inter-task separation and intra-task cohesion.

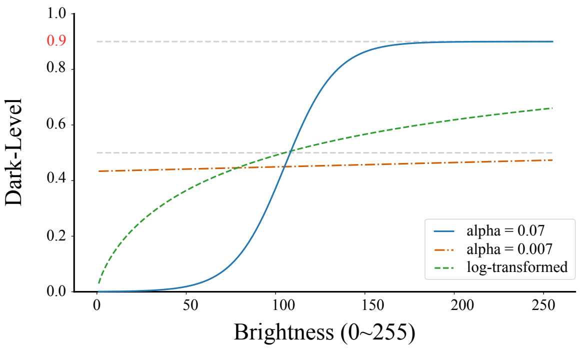
$\cos(\cdot,\cdot)$ denotes the mean cosine similarity across the first D' dimensions of the flattened feature vectors within each mini-batch. This formulation ensures that only the overlapping subspace between two representations is regularized, allowing the model to learn disentangled yet complementary task representations, and avoids the instability typically associated with gradient-level regularization.

3.4. Dark-Level Function: To quantify illuminance levels

To regulate the activation of the Rod and Cone blocks, it is necessary to estimate the darkness level of the current environment. However, simply using luminance as a proxy for darkness fails to account for the properties of rod cells, which do not become fully saturated and remain slightly responsive even under high-light conditions to detect contrast [34]. Therefore, it is essential to develop a more appropriate formulation for estimating the darkness level. The darkness level should be a continuous value in the range (0, 1) to represent the relative activation of photoreceptor cells. Under this condition, the darkness level can be interpreted as the probability of rod cell or cone cell activation. In deep learning frameworks, the sigmoid function is commonly used to convert logits into probability-like values. Thus, we formulate the darkness level using a sigmoid-based approach to reflect this probabilistic interpretation.

The formulation adopts a sigmoid-based structure to align with the probability-like representation (Equation 3).

Figure 3. Comparison of dark-level formulations. The green dotted line corresponds to the exponential-based formulation in Equation 7, mapped onto the 0–255 brightness range. The blue solid and orange dash-dotted lines represent the sigmoid-based formulation in Equation 3 with $\alpha = 0.07$ and $\alpha = 0.007$, respectively.



Since rod cells become saturated briefly when transitioning to high-light environments and subsequently recover to approximately 10–15% of their activation [34], the scaling factor τ is set to 0.9. This setting ensures that rod cells remain at least 10% active even under high luminance conditions.

$$\text{Dark-Level}(x) = \tau \times \frac{1}{1+e^{-\alpha(x-\beta)}} \tag{3}$$

As the formulation is scaled, the turning point at which the dark-level value equals 0.5 must be adjusted. This turning point represents the balance point between rod and cone activation and is commonly referred to as mesopic luminance [30]. Since the input image is gamma-corrected, its luminance must be interpreted in the RGB color space. Based on Equations 4–6, the mesopic luminance corresponds to a pixel value of approximately 107 when $\gamma = 2.2$, which equates to a relative luminance of 0.15.

$$\begin{aligned}\text{pixel value} &= \left(\frac{L}{L_{\text{max}}}\right)^{1/\gamma} \times 255 \quad (4)\\ &= (0.15)^{1/2.2} \times 255 \quad (5)\\ &\approx 107 \quad (6)\end{aligned}$$

To determine the optimal value of α, the formulation is derived from the relative cone activation weight W_{Cone} as proposed in [30]. The original expression (Equation 7) is transformed into a sigmoid-like form using the logarithmic identity $\left(\frac{M}{L}\right)^k = e^{k\cdot\ln(M/L)} = e^{k(\ln M - \ln L)}$.

$$
\begin{aligned}
W_{\text{Cone}} &= \frac{1}{1+\left(\frac{M}{L}\right)^k} && (7)\\
&= \frac{1}{1+e^{k(\ln M - \ln L)}} && (8)\\
&= \frac{1}{1+e^{k(\ln M - \ln(c\cdot x))}} && (9)\\
&= \frac{1}{1+e^{k(\ln(M/c) - \ln x)}} && (10)\\
&= \frac{1}{1+e^{-k(\ln x - \ln(M/c))}} && (11)
\end{aligned}
$$

Given the hypothesis that Equation 3 resembles Equation 11, the parameter α can be estimated by taking the derivative of each equation with respect to x. By equating the centers of activation, where both functions reach a value of 0.5, the relationship $\alpha(x-\beta) \approx k(\ln M - \ln(c\cdot x))$ is established. Assuming $K = 0.75$ as in [30] and $\beta = 107$ calculated in Equation 6, the value of α is approximated as 0.0071.

$$
\begin{aligned}
&\frac{d}{dx}\left(\frac{1}{1+e^{-k(\ln x - \ln(M/c))}}\right) && (12)\\
&= \frac{k/x \cdot e^{-k(\ln x - \ln(M/c))}}{\left(1+e^{-k(\ln x - \ln(M/c))}\right)^2} && (13)\\
&= \frac{k/x \cdot e^{-\alpha(x-\beta)}}{(1+e^{-\alpha(x-\beta)})^2} && (14)\\
&\approx \frac{d}{dx}\left(\frac{1}{1+e^{-\alpha(x-\beta)}}\right) && (15)\\
&= \frac{\alpha\cdot e^{-\alpha(x-\beta)}}{(1+e^{-\alpha(x-\beta)})^2} && (16)
\end{aligned}
$$

$$
\therefore \alpha = \frac{K}{b} = \frac{0.75}{107} \approx 0.0071 \tag{17}
$$

However, the dark-level function with $\alpha = 0.007$ exhibits a slope that is too shallow, as it fails to produce output values approaching 0 and 0.9 when the input pixel values tend toward the lower and upper extremes of the range, respectively (see Figure 3). To address this limitation, the value of α was scaled by a factor of 10, resulting in $\alpha = 0.07$. This practical adjustment is in line with strategies adopted in prior enhancement methods such as LIME [9], where regularization weights or gamma values are empirically tuned for perceptual quality. Based on this modification, the dataset was labeled using the adjusted formulation, and BRNet was trained on them. The final dark-level formulation is given as follows:

$$
\text{Dark-Level}(x) = 0.9 \times \frac{1}{1+e^{-0.07(x-107)}} \tag{18}
$$

4. Experiments

4.1. Experimental Setup

4.1.1 Datasets

We evaluate our method on three public datasets with varying illumination conditions:

WIDER FACE [42] is a large-scale face detection benchmark containing 32,203 images and 393,703 labeled faces under diverse conditions, including variations in pose, scale, and occlusion. It serves as the source domain for general face detection.

DARK FACE [2] consists of 6,000 real-world nighttime images designed for face detection under extremely low-light conditions. We use it as the target domain for both zero-shot and supervised evaluations.

ExDark [24] contains 7,363 images captured in various dark environments such as streets, tunnels, and indoor scenes. Although not limited to face detection, it provides object-level annotations including a face class. We use ExDark to assess model robustness in diverse low-light scenarios.

Together, these datasets provide a comprehensive benchmark for evaluating cross-domain generalization and low-light detection performance.

4.1.2 Metric

To evaluate detection performance and domain robustness, we use the widely adopted object detection metric: mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5, denoted as mAP@0.5.

Average Precision (AP) measures the area under the precision-recall curve for a given class. A prediction is considered correct if the predicted bounding box has an IoU of at least 0.5 with the ground truth box. Precision and recall are computed by comparing predicted boxes to the annotated ground truth. The IoU is defined as:

$$
IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}
$$

mAP@0.5 is obtained by averaging the AP across all object categories (in our case, primarily faces) with the IoU threshold fixed at 0.5. This metric reflects both localization accuracy and classification confidence, and is particularly important under low-light conditions where object boundaries may be unclear.

4.1.3 Evaluation Settings

We evaluate BRNet under two settings:

Zero-shot: The model is trained only on the WIDER FACE dataset and evaluated on a subset of 600 images randomly sampled from the DARK FACE validation set, since the official test set is no longer available for submission. This setting measures cross-domain generalization from bright to dark conditions, simulating practical scenarios where labeled low-light data is unavailable.

Fully-supervised: The model is fine-tuned on the labeled training split of DARK FACE and evaluated on the same 600-image validation subset. This setting reflects upper-bound performance when full supervision in the target domain is available.

4.2. Implementation Details

We trained BRNet using various data augmentation techniques, including photometric distortion, spatial transformations, random cropping, horizontal flipping, and interpolation-based resizing, with the augmentation strength set to 0.5. The model was optimized using SGD with a momentum of 0.9, weight decay of $5e^{-4}$, and an initial learning rate of $5e^{-4}$, which was linearly decayed throughout training. As shown in Section 3.4. The dark-level function parameters α, β, and τ were set to 0.06, 110, and 0.9, respectively. All experiments were conducted on an NVIDIA RTX A6000 GPU (48GB) with a batch size of 4, and the training was run for 100 epochs.

4.3. Results on DARK FACE

To evaluate the effectiveness of our proposed BRNet in low-light conditions, we conducted extensive experiments on the DARK FACE benchmark under three different settings: zero-shot adaptation, fully supervised learning with labeled dark data, and pretrained and tuned, where low-light preprocessing and pretrained models are used. The detailed results are presented in Table 1.

Table 1. Performance comparison under various training and adaptation settings on the DARK FACE validation set.

Category	Method	mAP(%)
WIDER FACE → DARK FACE validation set using DSFD		
Zero-shot Adaptation	CIConv [18]	18.4
	Sim-MinMax [25]	25.7
	DAI-Net [6]	28.0
	BRNet (Ours)	**-**
Fully Supervised	Fine-tuned DSFD [19]	46.0
	Fine-tuned DAI-Net [6]	52.9
	Fine-tuned BRNet	**-**
COCO → DARK FACE validation set using YOLOv3		
Pretrained and Tuning	YOLO$_N$	48.3
	YOLO$_N$+MBLLEN [26]	51.6
	YOLO$_N$+KIND [46]	51.6
	YOLO$_N$+Zero-DCE [8]	54.2
	YOLO$_L$	54.0
	MAET [3]	55.8
	DAI-Net [6]	57.0
	BRNet (Ours)	**-**

4.3.1 Zero-shot Adaptation

In the zero-shot adaptation setting, models are trained only on the WIDER FACE dataset and evaluated on the DARK FACE test split, so it never sees genuine dark images during training. Instead, we employ the non-deep method, DARK ISP, to generate synthetic dark counterparts of original images. This setting assesses the generalization capacity under domain shift, well-lit to low-light conditions. As seen in the Table 1, CIConv [18], Sim-MinMax [25], DAI-Net [6] achieved performance of 18.4, 25.7, 28.0 mAP, respectively. In contrast, BRNet(ours) demonstrated remarkable zero-shot generalization, achieving ? mAP, outperforming DAI-Net by ?. This result proves that our model effectively bridges the domain gap between high-visibility and low-visibility environments via a combination of dynamic luminance-sensitive encoding and structured architectural components. In particular, the proposed weighted combination of the Rod and Cone path using dark-level allows the network to contribute significantly better performance, as it allows the model to selectively emphasize high-frequency or contrast-preserving features based on the image's darkness.

4.3.2 Fully Supervised

In the fully supervised setting, models are fine-tuned using annotated samples from the DARK FACE training dataset. This configuration reflects performance under optimal data availability for the target domain. A baseline fine-tuned DSFD detector records 46.0 mAP, and fine-tuned DAI-Net reaches 52.9 mAP. After fine-tuning BRNet on the DARK FACE dataset, our model achieves ? mAP, beating DSFD by ? mAP points and DAI-Net by ? mAP points. In this case, our model shows improved results compared with its zero-shot setting(? mAP), which demonstrates that BRNet not only exhibits strong generalization without supervision, but also efficiently incorporates target-domain cues when domain labels are available.

4.3.3 Pretrained and Tuning

In the pretrained and tuning setting, we evaluate models on the DARK FACE validation set using detectors pretrained on COCO. Here, YOLO$_{\mathrm{N}}$ refers to a standard YOLOv3

Table 2. Comparison with state of the art on ExDark.

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	**Total**
YOLO$_N$	71.8	64.5	63.9	81.6	76.8	55.4	49.7	56.8	63.8	61.8	65.7	40.5	62.7
+KinD [46]	73.4	68.1	65.5	86.2	78.3	63.0	56.9	62.7	68.2	67.1	69.6	48.2	67.3
+Zero-DCE [8]	79.5	71.3	70.4	89.0	80.7	68.4	65.7	68.6	75.4	67.2	76.2	51.1	72.0
YOLO$_L$	78.2	70.8	72.3	88.1	80.7	67.9	62.4	70.5	74.8	69.4	75.8	50.9	71.6
MAET [3]	81.3	71.6	74.5	89.7	82.1	69.5	65.5	72.6	75.4	72.7	77.4	53.3	74.0
DAI-Net [6]	83.8	75.8	75.1	94.2	84.1	74.9	73.1	79.2	82.2	76.4	80.7	59.8	78.3
BRNet(Ours)	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 3. Ablation study of proposed components (s-ort loss, gain module, tapetum module, cone block) on DARK FACE.

Method	**Gain**	**Tapetum**	**Cone**	$\mathcal{L}_{\text{s-ort}}$	**mAP(%)**
A	–	–	–	–	-
B	✓	–	–	–	-
C	✓	✓	–	–	-
D	✓	✓	✓	–	-
E	✓	✓	✓	✓	-
F	–	✓	✓	✓	-
G	–	–	✓	✓	-

model trained on normal-light images, while YOLO$_\text{L}$ employs a larger backbone trained on synthetic low-light images generated via Dark ISP [6].

As shown in Table 1, YOLO$_\text{N}$ achieves 48.3 mAP, while image enhancement methods such as MBLLEN, KIND, and Zero-DCE improve performance up to 54.2 mAP. YOLO$_\text{L}$ achieves 54.0 mAP, showing limited gains from increased capacity. DAI-Net [6], which leverages domain adaptation and synthetic low-light training, achieves the highest prior score of 57.0 mAP.

In comparison, BRNet achieves ? mAP, surpassing all previous approaches. Without relying on enhancement or synthetic data, BRNet achieves robust performance through its biologically grounded dual-path design and task-aware regularization.

4.4. Cross-Evaluation on ExDark

To evaluate the generalization ability of BRNet under diverse low-light conditions, we conduct cross-dataset experiments on the ExDark benchmark. Unlike DARK FACE, which focuses primarily on pedestrian detection in urban night scenes, ExDark includes 7,363 low-light images spanning 12 object categories and varying illumination levels from extremely dark scenes. This setup enables evaluation of the model's robustness to unseen object types and lighting variations without additional fine-tuning.

As shown in Table 2, previous methods such as YOLO$_N$ with Zero-DCE, MAET, and DAI-Net achieved average mAP scores of 72.0, 74.0, and 78.3, respectively. In comparison, BRNet achieves an mAP of **?**, outperforming prior approaches and demonstrating superior cross-domain generalization.

We attribute this performance to BRNet's biologically inspired Photo Receptor block and the use of semi-orthogonal loss. The learned dark-level–aware weighting of rod and cone pathways enables the model to extract illumination-invariant features, maintaining high detection accuracy even under extreme nighttime conditions and category shifts.

4.5. Ablation Study

To evaluate the individual contribution of each proposed component in BRNet, we conduct an ablation study under a zero-shot adaptation setting from WIDER FACE to DARK FACE. The components under investigation include the Gain module, Tapetum module, Cone block, and Semi Orthogonal Loss, which are described in Sections 3.2 and 3.3. We construct multiple variants of BRNet by incrementally enabling these components and report their mAP scores in Table 3.

Starting from the baseline (variant A), which excludes all proposed modules, we observe that introducing the Gain module (variant B) improves the model's sensitivity to local contrast in low-light regions by enhancing high-frequency features. Adding the Tapetum module (variant C) further boosts performance, as reflectance-aware signal amplification is particularly effective under extremely dark conditions. Incorporating the Cone block (variant D) enhances the model's ability to capture spatial semantics in relatively brighter areas, improving robustness across a broad range of illumination levels.

The addition of the Semi Orthogonal Loss (variant E) leads to the highest performance, as it mitigates task interference between the main detection and auxiliary prediction branches by selectively disentangling shared representations. This balance between feature separation and task synergy enables the model to benefit from multi-task learning without sacrificing detection performance.

Finally, variants F and G remove the Gain or Tapetum modules from the full model. This results in noticeable performance degradation, confirming that both modules are essential for extracting meaningful signals in low-illumination environments. These results demonstrate that each component contributes uniquely and complementarily to BRNet's performance. Their combined effect is key to achieving robust generalization under adverse lighting conditions.

5. Conclusion

We proposed BRNet, a low-light object detector inspired by biological vision. The architecture mimics the structure of retinal photoreceptors. It separates processing into Rod and Cone pathways, regulated by dark-level estimation. The Rod block captures fine contrast and reflectance cues through the Gain and Tapetum modules. Deformable convolution further enhances the model's ability to handle spatial variations in low-light scenes.

We also introduced a Semi Orthogonal Loss to reduce interference between tasks. This regularization preserves task-specific representations while allowing synergy across branches. Experiments on DARK FACE and ExDark show that each module improves performance. Together, they enable strong generalization without retraining. Our results highlight the effectiveness of combining biologically grounded design with structured feature disentanglement for robust low-light detection.

References

[1] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12504–12513, 2023. 2

[2] Wenhan Yang Jiaying Liu Chen Wei, Wenjing Wang. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018. 6

[3] 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*, pages 2553–2562, 2021. 2, 3, 4, 7, 8

[4] Christine A Curcio, Kenneth R Sloan, Robert E Kalina, and Anita E Hendrickson. Human photoreceptor topography. *Journal of comparative neurology*, 292(4):497–523, 1990. 3

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 4

[6] 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*, pages 12666–12676, 2024. 2, 7, 8

[7] Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas K Sharma, Tom Hartvigsen, and Marzyeh Ghassemi. Bendvlm: Test-time debiasing of vision-language embeddings. *Advances in Neural Information Processing Systems*, 37:62480–62502, 2024. 4

[8] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020. 1, 2, 7, 8

[9] Xiaojie Guo. Lime: A method for low-light image enhancement. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 87–91, 2016. 6

[10] 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*, pages 6725–6735, 2023. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[12] 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. 2

[13] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst technical report, 2010. 2

[14] 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

[15] Yeying Jin, Wenhan Yang, and Robby T Tan. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In *European Conference on Computer Vision*, pages 404–421. Springer, 2022. 2

[16] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14992–15001, 2021. 4

[17] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 2

[18] 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*, pages 4399–4409, 2021. 1, 2, 7

[19] 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

[20] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022. 4

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 2

[22] 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. 2

[23] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019. 2

[24] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. 2, 6

[25] 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. 7

[26] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *Bmvc*, volume 220, page 4. Northumbria University, 2018. 2, 7

[27] 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*, pages 5637–5646, 2022. 2

[28] 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 *Proceedings of the 30th ACM international conference on multimedia*, pages 2070–2078, 2022. 2

[29] 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. 2

[30] Sabine Raphael and Donald IA MacLeod. Mesopic luminance assessed with minimum motion photometry. *Journal of Vision*, 11(9):14–14, 2011. 5, 6

[31] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[32] Liang Shen, Zihan Yue, Fan Feng, Quan Chen, Shihao Liu, and Jie Ma. Msr-net: Low-light image enhancement using deep convolutional network. *arXiv preprint arXiv:1711.02488*, 2017. 2

[33] Stefan Smeu, Elisabeta Oneata, and Dan Oneata. Declip: Decoding clip representations for deepfake localization. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 149–159. IEEE, 2025. 4

[34] Alexandra Tikidji-Hamburyan, Katja Reinhard, Riccardo Storchi, Johannes Dietter, Hartwig Seitter, Katherine E Davis, Saad Idrees, Marion Mutter, Lauren Walmsley, Robert A Bedford, et al. Rods progressively escape saturation to drive visual responses in daylight conditions. *Nature communications*, 8(1):1813, 2017. 5

[35] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*, pages 1785–1797, 2021. 4

[36] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6849–6857, 2019. 2

[37] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14408–14419, 2023. 3, 4

[38] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 1, 2

[39] 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*, pages 5901–5910, 2022. 2

[40] Xinpeng Xie, Jiawei Chen, Yuexiang Li, Linlin Shen, Kai Ma, and Yefeng Zheng. Self-supervised cyclegan for object-preserving image-to-image domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 498–513. Springer, 2020. 1, 2

[41] 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*, pages 2154–2162, 2022. 2

[42] 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. 2, 6

[43] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. 2

[44] Yongxin Yang and Timothy Hospedales. Deep multi-task representation learning: A tensor factorisation approach. *arXiv preprint arXiv:1605.06391*, 2016. 2

[45] Xiangchen Yin, Zhenda Yu, Zetao Fei, Wenjun Lv, and Xin Gao. Pe-yolo: Pyramid enhancement network for dark object detection. In *International conference on artificial neural networks*, pages 163–174. Springer, 2023. 2

[46] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1632–1640, 2019. 2, 7, 8