Multitask AET with Orthogonal Tangent Regularity for Dark Object Detection

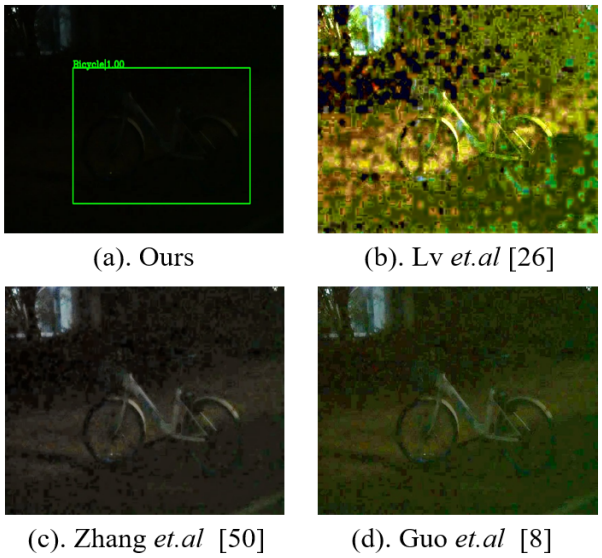
Ziteng Cui[1], Guo-Jun Qi[2], Lin Gu[3,4*], Shaodi You[5], Zenghui Zhang[1], Tatsuya Harada[4,3]
[1]Shanghai Jiao Tong University [2], Seattle Research Center, Innopeak Technology
[3]RIKEN AIP, [4] The University of Tokyo, [5] University of Amsterdam
cuiziteng@sjtu.edu.cn, guojunq@gmail.com, lin.gu@riken.jp, s.you@uva.nl, zenghui.zhang@sjtu.edu.cn, harada@mi.t.u-tokyo.ac.jp

Abstract

Dark environment becomes a challenge for computer vision algorithms owing to insufficient photons and undesirable noise. To enhance object detection in a dark environment, we propose a novel multitask auto encoding transformation (MAET) model which is able to explore the intrinsic pattern behind illumination translation. In a self-supervision manner, the MAET learns the intrinsic visual structure by encoding and decoding the realistic illumination-degrading transformation considering the physical noise model and image signal processing (ISP). Based on this representation, we achieve the object detection task by decoding the bounding box coordinates and classes. To avoid the over-entanglement of two tasks, our MAET disentangles the object and degrading features by imposing an orthogonal tangent regularity. This forms a parametric manifold along which multitask predictions can be geometrically formulated by maximizing the orthogonality between the tangents along the outputs of respective tasks. Our framework can be implemented based on the mainstream object detection architecture and directly trained end-to-end using normal target detection datasets, such as VOC and COCO. We have achieved the state-of-the-art performance using synthetic and real-world datasets. Codes will be released at https://github.com/cuiziteng/MAET.

1. Introduction

Low-illumination environment poses significant challenges in computer vision. Computational photography community has proposed many human-vision-oriented algorithms to recover normal-lit images [26, 46, 28, 27, 4, 20, 51, 12, 8]. Unfortunately, the restored image does not necessarily benefit the high-level visual understanding tasks. As the enhancement/restoration approaches are optimized for human visual perception, they may generate artifacts (see Fig. 1 for an example), which are misleading for consequent vision tasks. Another line of research focuses on the robustness of specific high-level vision algorithms. They either train models on a large volume of real-world data [31, 25, 48] or rely on carefully designed task-related features [29, 17].



(a). Ours (b). Lv *et.al* [26]
(c). Zhang *et.al* [50] (d). Guo *et.al* [8]

Figure 1. Detection and enhancement results of image taken by Sony DSC-RX100M7 camera at night with 0.1s exposure time and 3200 ISO. (a) is detection result on the original image by MAET (YOLOv3) and (b), (c), (d) are the enhanced images by Lv *et al.* [28], Zhang *et al.* [51], Guo *et al.* [8] respectively, on which YOLOv3 failed to make detection.

However, existing methods suffer from two major inconsistencies: *target inconsistency and data inconsistency* (in the existing research). *Target inconsistency* refers to the fact that most methods focus on their own target, either human vision or machine vision. Each line follows their routes separately without benefiting each other under a general framework.

In the meantime, *data inconsistency* complicates the assumption that the training data should resemble the one used

*Corresponding author.

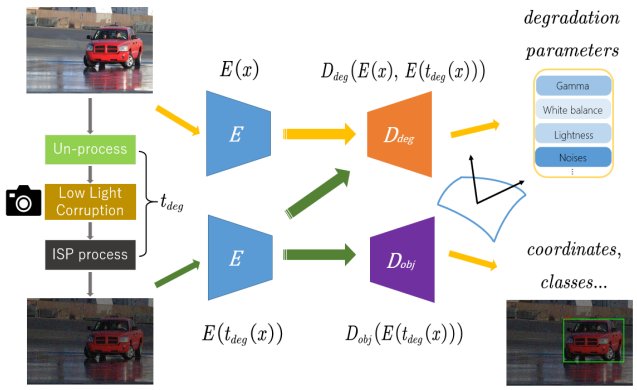


Figure 2. Structure of the multitask autoencoding transformations (MAET) framework.

for evaluation. For example, pre-trained object detection models are usually trained on clear and normal lit images. To adapt to the poor light condition, they rely on the augmented dark images to fine-tune the models without exploring the intrinsic structure under the illumination variance. Just like *Happy families are all alike; every unhappy family is unhappy in its own way*, even with the existing datasets [6, 23, 16], the varied distribution of real-world conditions can hardly be covered by the training set.

Here, we aim to bridge above two gaps under a unified framework. As illustrated in Fig. 2, the normal lit image can be parametrically transformed (t_{deg}) into their degraded low-illumination counterparts. Based on this transformation, we propose a novel multitask autoencoding transformation (MAET) to extract the transformation equivariant convolutional features for object detection in dark images. We train the MAET based on two tasks: (1) to learn the intrinsic representation by decoding the low-illumination-degrading transformation based on unlabeled data and (2) to decode object position and categories based on labeled data. As shown in Fig. 2, we train our MAET to encode the pair of normal lit and low-light images with siamese encoder E and decode its degrading parameters, such as noise level, gamma correction, and white balance gains, using decoder D_{deg}. This allows our model to capture the intrinsic visual structure that is equivariant to illumination variance. Compared with [26, 42, 20, 28, 41], who conducted over-simplified synthesis, we design our degradation model considering the physical noise model of sensors and image signal processing (ISP). Then, we perform the object detection task by decoding the bounding box coordinates and classes with the decoder D_{obj} based on the representation encoded by E (Fig. 2).

Although MAET regularizes network training by predicting low-light degrading parameters, the joint training of object detection and transformation decoding are over-entangled through a shared backbone network. While this improves the detection of dark objects using MAET regularity, it may also risk overfitting the object-level representation into self-supervisory imaging signals. To this end, we propose to disentangle the object detection and transformation decoding tasks by imposing an orthogonal tangent regularity. It assumes that the multivariate outputs of above two tasks form a parametric manifold, and disentangling the multitask outputs along the manifold can be geometrically formulated by maximizing the orthogonality among the tangents along the output of different tasks. The framework can be directly trained end-to-end using standard target detection datasets, such as COCO [23] and VOC [6], and make it detect low-light images. Although we consider YOLOv3 [39] for illustration, the proposed MAET is a general framework that can be easily applied to other mainstream object detectors, *e.g.*, [40, 22, 54].

Our contributions to this study are as follows:

- By exploring physical noise models of sensors and the ISP pipeline, we leverage a novel MAET framework to encode the intrinsic structure, which can decode low-light-degrading transformation. Then, we perform the object detection by decoding bounding box coordinates and categories based on this robust representation. Our MAET framework is compatible with mainstream object architectures.
- Moreover, we present the disentangling of multitask outputs to avoid the overfitting of the learned object-detection features into the self-supervisory degrading parameters. This can be naturally performed from a geometric perspective by maximizing the orthogonality along the tangents corresponding to the output of different tasks.
- Based on comprehensive evaluation and compared with other methods, our method shows superior performance pertaining to low-light object detection tasks.

2. Related Work

2.1. Low Illumination Datasets

Several datasets have been proposed for the low-light object detection task: Neumann *et al.* [31] proposed NightOwls dataset for pedestrians detection in the night. Nada *et al.* [30] collected an unconstrained face detection dataset (UFDD) considering various adverse conditions, such as rain, snow, haze, and low illumination. In recent times, the UG^2+ challenge [48] has included several tracks for vision tasks under different poor visibility environments. Among them, DARK FACE dataset with 10,000 images (includes 6,000 labeled and 4,000 unlabeled images). For the multi-class dark object detection task, Loh *et al.* [25] proposed exclusively dark (ExDark) dataset, which includes 7363 images with 12 object categories.

2.2. Low-Light Vision

2.2.1 Enhancement and Restoration Methods

Low-light vision tasks focus on the human visual experience by restoring details and correcting the color shift. Early attempts are either Retinex theory based approaches [18, 13, 9] or histogram equalization (HE) based approaches [44, 19]. Nowadays, with the development of deep learning, CNN based methods [26, 46, 28, 27, 20, 51, 8] and GAN based methods [15, 12] have achieved a significant improvement in this task. Like Wei *et al.* [46] combined the Retinex theory [18] with deep network for low-light image enhancement. Jiang *et al.* [12] used an unsupervised GAN to solve this problem. Very recently, Guo *et al.* [8] proposed a self-supervised method, which could learn without normal light images.

2.2.2 High-Level Task

To adopt the high-level task for a dark environment, a straightforward strategy is casting the aforementioned enhancement methods as a post-processing step [53, 8]. Other ones rely on augmented real-world data [31, 25, 48, 47] or some oversimplified synthetic data [52, 41].Recent real noisy image benchmarks [2, 35] show that sometimes hand-crafted algorithms may even outperform deep learning models. To combine the strength of computational photograhy, we develop a framework with transformation-equivariant representation learning.

2.3. Transformation-Equivariant Representation Learning

Several self-supervised representation learning methods have been proposed to learn image features either through solving Jigsaw Puzzles [33] or impainting the missing region of an image [34]. Recently, a series of auto-encoding transformations (AETs), such as AET [50], AVT [37], EnAET [45], have demonstrated state-of-the-art performances for several self-supervised tasks. As the AET is flexible and not restricted to any specific convolutional structure, we extend it to our multitask AET for object detection in dark images.

3. Multitask Autoencoding Transformation (MAET)

In this section, we first briefly introduce autoencoding transformation (AET) [50], based on which we propose multitask AET (MAET). Then, we discuss the ISP pipeline in camera to design the degrading transformations to be leveraged by our MAET. Finally, we explain the MAET architecture and training and testing details.

3.1. Background: From AET to MAET

AET [50] learns representative latent features that decode or recover the parameterized transformation from the original image (x) and the transformed counterpart ($t(x)$) based on transformation t:

$$x \xrightarrow{\mathcal{T}} t(x). \tag{1}$$

The AET comprises a siamese representation encoder (E) and a transformation decoder (D). The encoder E extracts features from x and its transformation $t(x)$, which should capture intrinsic visual structures to explain the transformation t (*e.g.*, the low-illumination degrading transformations in the next section). Then, the decoder D uses the encoded $E(x)$ and $E(t(x))$ to decode the estimation $\hat{t}$ for t:

$$\hat{t} = D_\phi[E(x), E(t(x))]. \tag{2}$$

The AET, specifically the representation encoder E and transformation decoder D, can be trained by minimizing the deviation loss ℓ of the original transform t and the predicted result $\hat{t}$:

$$\mathcal{L}_{aet} \triangleq \sum_k \ell_k(\hat{t_k}, t_k), \tag{3}$$

where ℓ_k denotes type k transformation loss computed using the mean-squared error (MSE) loss between the predicted transformation $\hat{t_k}$ and ground truth transformation t_k.

3.2. Multi-Task AET with Orthogonal Regularity

In this study, we further extend the AET to the MAET by simultaneously solving multiple tasks. As illustrated in Fig. 2, the proposed MAET model consists of two parts: a representation encoder (E) and multi-task decoders. For the task of illumination-degrading transformation t_{deg}, we use decoder D_{deg} to decode the degrading parameters. The task of object detection is realized by the decoder D_{obj} to predict the bounding box location and object categories directly from illumination-degenerated images. Although the two tasks are correlated, their outputs reflect very different aspects of input images: the illumination conditions for D_{deg} and the object locations and categories for D_{obj}. This suggests that an orthogonal regularity can be imposed to decouple the unnecessary interdependence between the outputs of different tasks.

To this end, the orthogonal objective of the proposed MAET is to minimize the absolute value of cosine similarity below:

$$\mathcal{L}_{ort} \triangleq \sum_{k,l} |\cos\theta_{k,l}| = \sum_{k,l} \frac{\left|\left[\frac{\partial E}{\partial D^k_{deg}}\right]^T \cdot \left[\frac{\partial E}{\partial D^l_{obj}}\right]\right|}{\left\|\frac{\partial E}{\partial D^k_{deg}}\right\| \cdot \left\|\frac{\partial E}{\partial D^l_{obj}}\right\|}, \tag{4}$$

where $\frac{\partial E}{\partial D^k_{deg}}$ and $\frac{\partial E}{\partial D^l_{obj}}$ are the tangents of the representation manifold formed by the encoder E along the kth and

lth output coordinates of the illumination-degrading transformation and object detection tasks, respectively. In other words, these two tangents depict the directions along which the representation moves with the change of the decoder outputs D^k_{deg} and D^l_{obj}, respectively.

Minimizing the absolute value of the cosine similarity will push the two tangents as orthogonal as possible. Based on the geometric point of view, this will disentangle the two tasks so that the change of the predicted coordinates for one task will have a minimal impact on the coordinates for the other task. In Sections 3.3, we will discuss the details about how to define the low-illumination-degrading transformation. The idea of imposing orthogonality between tasks was explored in literature [43, 24, 49]. However, here we implement it in the context of AET, where the orthogonal directions are defined in terms of decoder tangents along the encoder-induced manifold, which differs from the previous works.

Therefore, the total loss for our low-light object detection consists of three parts: degradation transformation loss $\mathcal{L}_{deg}$, object detection loss $\mathcal{L}_{obj}$ and orthogonal regularity loss $\mathcal{L}_{ort}$ (cf. Eq. (4)), the total loss used for training can be represented as

$$\mathcal{L}_{total} = \mathcal{L}_{ort} + \omega_1 \cdot \mathcal{L}_{obj} + \omega_2 \cdot \mathcal{L}_{deg}. \tag{5}$$

The object detection loss $\mathcal{L}_{obj}$ is specific for different object detectors [40, 54, 39]. In this experiments, $\mathcal{L}_{obj}$ is the loss function of YOLOv3 [39], which includes location loss, classification loss and confidence loss. The degradation transformation loss $\mathcal{L}_{deg}$ is the AET loss (cf. Eq. (3)) with the low-illumination degrading transformation t_{deg}, and ω_1 and ω_2 are the fixed balancing hyper-parameters.

3.3. Low-Illumination Degrading Transformations

Given a normal lit noise-free image x, we aim to design a low-illumination-degrading transformation t_{deg} to transform x into a dark image $t_{deg}(x)$ that matches the real photo captured under low-light conditions, *i.e.*, by turning off the light. Most of existing methods conduct an over-simplified synthesis, *e.g.*, invert gamma correction (sometimes with additive mixed Gaussian noise) [26, 28, 52] or retinex theory based synthesis method [20] [1]. The ignorance of the physics of sensors and on-chip image signal processing (ISP) makes these methods generalize poorly to real-world dark images. Here, we first systematically describe the ISP pipeline between the sensor measurement system and the final photo. Based on this pipeline, we parametrically model the low light-degrading transformation t_{deg}.

[1]The contrast experiments with these synthesis methods have been given in our supplementary materials Appendix.B.1.

3.3.1 Image Signal Processing (ISP) Pipeline

The camera is designed to render the photo to be as pleasant and accurate as possible based on the perspective of a human eye. For this reason, the RAW data captured by the camera sensor requires ISP (several steps) before becoming the final photo. Much research has been done to simulate this ISP process [38, 10, 11, 32, 21]. For example, Karaimer and Brown [14] step-by-step detailed the ISP process and showed its high potential pertaining to computer vision.

We adopt a simplified ISP and its unprocessing procedure from [3] (Fig. 3). Particularly, we ignore several steps including the demosaicing process [1]. Although these processes are important for precise ISP algorithm, most images on the Internet are of various sources and do not follow the perfect ISP procedure. We ignore these steps for the trade-off between precision and generability. We have made a detailed analysis of the demosaicing's influence in supplementary materials Appendix.B.2. Next, we introduce our ISP process in detail.

Quantization is the analog voltage signal step that quantizes the analog measurement x into discrete codes y_{quan} using an analog-to-digital converter (ADC). The quantization step maps a range of analog voltages to a single value and generates a uniformly distributed quantization noise. To simulate the quantization step, the quantization noise x_{quan} related to B bits has been added. In our degrading model, B is randomly chosen from 12, 14, and 16 bits.

$$\begin{aligned} x_{quan} &\sim U(-\frac{1}{2B}, \frac{1}{2B}) \\ y_{quan} &= x + x_{quan}. \end{aligned} \tag{6}$$

White Balance simulates the color constancy of human vision system (HVS) to map "white" colors with the white object [38]. The captured image is the product of the color of light and material reflectance. The white-balance step in the camera pipeline estimates and adjusts the red channel gain g_r and blue channel gain g_b to make image appearing to be lit under "neutral" illumination.

$$\begin{bmatrix} y_r \\ y_g \\ y_b \end{bmatrix} = \begin{bmatrix} g_r & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & g_b \end{bmatrix} \cdot \begin{bmatrix} x_r \\ x_g \\ x_b \end{bmatrix} \tag{7}$$

Based on [35, 3], g_r is randomly chosen from $(1.9, 2.4)$, and g_b is randomly chosen from $(1.5, 1.9)$; both follow an uniform distribution and are independent of each other. The inverse process considers the reciprocal of the red and blue gains $1/g$.

Color Space Transformation converts the white-balanced signal from camera internal color space $cRGB$ to $sRGB$ color space. This step is essential in ISP pipeline as camera color space are not identical to the sRGB space

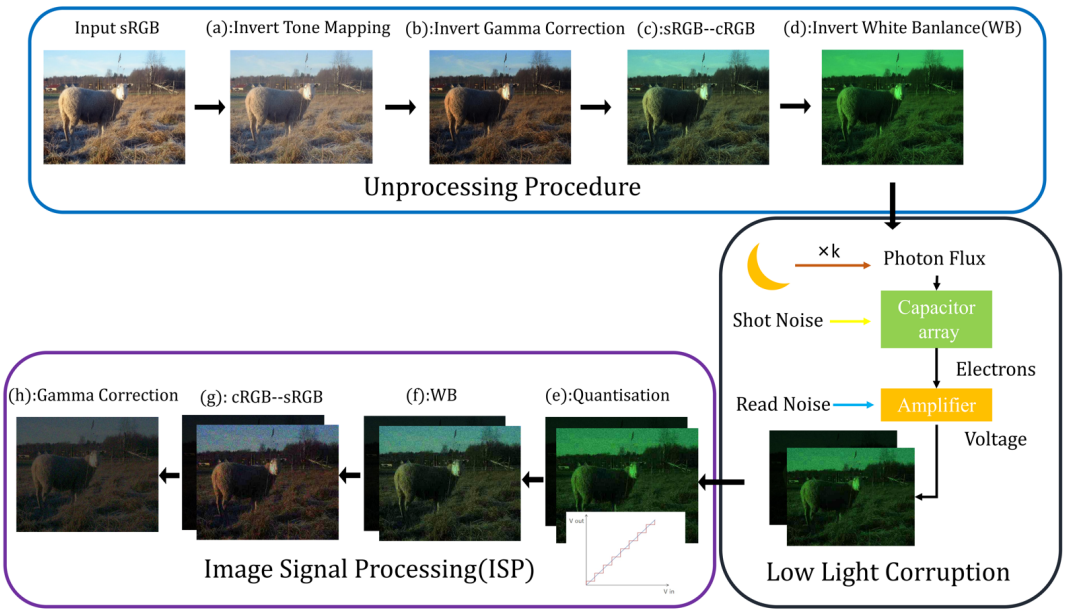


Figure 3. General view of the low-illumination-degrading pipeline, an sRGB "sheep" image from *PASCAL VOC2007 dataset* [6] passed by unprocessing procedure, low-light corruption, and image signal processing(ISP) process to get the final degraded low-light counterpart

[38, 14]. The converted signal y_{sRGB} can be obtained with a 3×3 color correction matrix (CCM) M_{ccm}:

$$y_{sRGB} = M_{ccm} \cdot y_{cRGB}, \tag{8}$$

the inversion of this process is:

$$y_{cRGB} = M_{ccm}^{-1} \cdot y_{sRGB}. \tag{9}$$

Gamma Correction has also been widely used in the ISP pipeline for the non-linearity of humans perception on dark areas [36]. Here we use the standard gamma curve [35] as:

$$y_{gamma} = max(x, \epsilon)^{\frac{1}{\gamma}} \tag{10}$$

and its' invert process is:

$$y_{invert\ gamma} = max(x, \epsilon)^{\gamma}. \tag{11}$$

The gamma curve parameter γ could be randomly sampled from an uniform distribution $\gamma \sim U(2,\ 3.5)$ and ϵ is a very small value ($\epsilon = 1e^{-5}$) to prevent numerical instability during training.

Tone Mapping aims to match the "characteristic curve" of film. For the sake of computational complexity, we perform a "smoothstep" curve [3] as

$$y_{tone} = 3x^2 - 2x^3 \tag{12}$$

and we could also perform the inverse with:

$$y_{invert\ tone} = \frac{1}{2} - sin(\frac{sin^{-1}(1-2x)}{3}). \tag{13}$$

Figure 4. Examples of our degrading transformation on SID dataset [4]. The long-exposure RAW images and their groundtruth short-exposure RAW images are transformed into the sRGB format using *Adobe Lightroom*, separately shown in the first and second columns. The third column shows the images generated from our pipeline.

3.3.2 Degrading Transformation Model

After defining each step of our ISP pipeline, we can present our low-illumination-degradation transform t_{deg} that synthesizes realistic dark light image $t_{deg}(x)$ based on its normal light counterparts x. At first, as shown in Fig. 3, we have to use an inverse processing procedure [3] to transform the normal lit image x into sensor measurement or RAW data. Then, we linearly attenuate the RAW image and corrupt it with shot and read noise. Finally, we continue applying the pipeline to turn the low-lit sensor measurement to the photo $t_{deg}(x)$.

Unprocessing Procedure: Based on [3], the unprocessing part aims to translate the input $sRGB$ images into their

RAW format counterparts, which are linearly proportional to the captured photons. As shown in Fig. 3, we unprocess the input images by (a) invert tone mapping, (b) invert gamma correction, (c) transformation of image from $sRGB$ space to $cRGB$ space, and (d) invert white balancing, here we call (a), (b), (c), (d) together as $t_{unprocess}$. Based on these parts, we synthesize realistic RAW format images, and the resulting synthetic RAW image is used for low-light corruption process.

Low Light Corruption:

When light photons are projected through a lens on a capacitor cluster, considering the same exposure time, aperture size, and automatic gain control, each capacitor develops an electric charge corresponding to the lux of illumination of the environment.

Shot noise is a type of noise generated by the random arrival of photons in a camera, which is a fundamental limitation. As the time of photon arrival is governed by Poisson statistics, uncertainty in the number of photons collected during a given period is $\delta_s = \sqrt{S}$, where δ_s is the shot noise and S is the signal of the sensor.

Read noise occurs during the charge conversion of electrons into voltage in the output amplifier, which can be approximated using a Gaussian random variable with zero mean and fixed variance.

Shot and read noises are common in a camera imaging system; thus, we model the noisy measurement x_{noise} [7] on the sensor:

$$\begin{aligned} x_{noise} &\sim N(\mu = kx, \sigma^2 = \delta_r^2 + \delta_s kx) \\ y_{noise} &= kx + x_{noise}, \end{aligned} \tag{14}$$

where the true intensity of each pixel x from the unprocessing procedure. We linearly attenuate it with parameter k. To simulate different lighting conditions, the parameter of light intensity k is randomly chosen from a truncated Gaussian distribution, in range of $(0.01, 1.0)$, with mean 0.1 and variance 0.08. The parameter range of δ_r and δ_s follows [35], which is shown in Table 1.

ISP Pipeline: RAW images often pass through a series of transformations, before we see it in the RGB format; therefore, we apply RAW image processing after the low-light corruption process. Based on [3], our transformations are in the following order: (e) add quantization noise (f) white balancing, (g) from $cRGB$ to $sRGB$, and (h) gamma correction, we call (f), (g), (h) together as t_{ISP}.

Finally, we can obtain the degraded low-light images $t_{deg}(x)$ from the noise-free x, as it shown in Eq. 15. Some examples of the original images, generated images, and groundtruth are shown in Fig. 4. We summarize the parameters and their ranges involved in t_{deg} (Table 1):

$$t_{deg}(x) = t_{ISP}(k \cdot t_{unprocess}(x) + x_{noise} + x_{quan}). \tag{15}$$

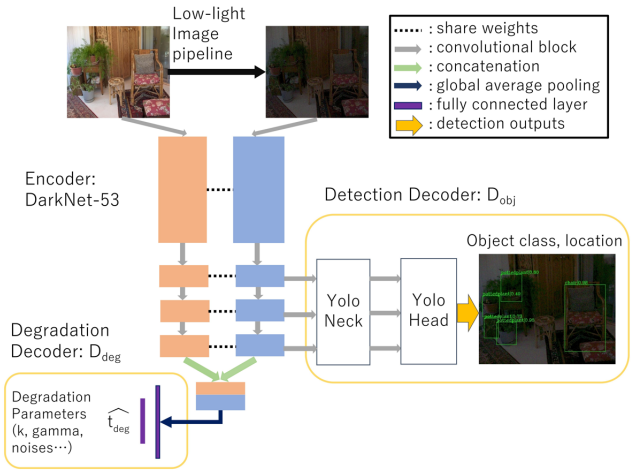


Figure 5. Architecture of the proposed MAET model base on YOLOv3 framework.

3.4. Architecture

The architecture of the proposed MAET is shown in Fig. 5. Our network comprises representation encoder E and decoder D. For illustration, we implement the MAET based on the architecture of YOLOv3 [39]. Moreover, this can be replaced by other mainstream detection frameworks, *e.g.* , [40], [22], [54].

E adopts a siamese structure with shared weights. During the training process, the normal lit image x is fed into the left path of E (denoted in orange), while its degraded counterparts $t_{deg}(x)$ go through the right path or dark path (denoted in blue). Here, the encoder adopts DarkNet-53 network [39] as the backbone.

As we solve two tasks, degrading transformation decoding and object detection tasks, decoder D can be divided into degrading transformation decoder D_{deg} and object detection decoder D_{obj}. The former focuses on decoding the parameters of low-light-degrading transformation (t_{deg}). The latter decodes the target information, *i.e.*, target class and location. As shown in Fig. 5, the encoded latent features $E(x)$ and $E(t_{deg}(x))$ are concatenated together and passed to decoder D_{deg} to estimate the corresponding degrading transformation t_{deg}. This self-supervision training helps the MAET learn the intrinsic visual structure under various illumination degrading transformations with unlabeled data. The object detection decoder D_{obj} only decodes the representation $E(t_{deg}(x))$ from the dark path (denoted in blue) to predict the parameters of object detection. In the testing time, we directly feed the low-light images to the dark path of the MAET encoder to decode the detection results: target categories and locations.

Step	Transformation	Range	Parameter(s) to Learn
Light Intensity	$f(x) = k \cdot x$	$k \sim N(\mu = 0.1, \sigma = 0.08)$ $0.01 \leq k \leq 1.0$	k
Shot Noise and Read Noise	$f(x) = x + N(\mu = x, \sigma^2 = \delta_r^2 + \delta_s x)$	$log\delta_s \sim U(-4, -2)$ $\frac{log\delta_r^2}{log\delta_s} \sim N(\mu = 2.18 log\delta_s + 0.12, \sigma = 0.26)$	-
Quantization	$f(x) = x + U(-\frac{1}{2B}, \frac{1}{2B})$	$B \in [12, 14, 16]$	$\frac{1}{B}$
White Balance	$f(x) = \begin{bmatrix} g_r & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & g_b \end{bmatrix} \cdot x$	$g_r \sim U(1.9, 2.4)$ $g_b \sim U(1.5, 1.9)$	$\frac{1}{g_r}, \frac{1}{g_b}$
Color Correction $cRGB \rightarrow uRGB \rightarrow sRGB$	$f(x) = M_{cu} \cdot M_{us} \cdot x$	Mixture of four color correction matrices (CCMs): Sony A7R, Olympus E-M10, Sony RX100 IV, Huawei Nexus 6P in [3]	-
Tone Mapping	$f(x) = 3x^2 - 2x^3$	-	-
Gamma Correction	$f(x) = max(x, \epsilon)^{\frac{1}{\gamma}}$, $\epsilon = 1e^{-5}$	$\gamma \sim U(2,\ 3.5)$	$\frac{1}{\gamma}$

Table 1. Details of low-illumination-degrading transformation parameters, the first column denotes the names of the transformations, the second column denotes the transformation process, the third column is the parameter range, and the last line denotes the parameters to be predicted in our MAET model's degrading transformation decoder.

	training set	testing set	pre-process	VOC(AP_{50})	COCO(AP)	COCO(AP_{50})	COCO(AP_{75})	COCO(AP_S)	COCO(AP_M)	COCO(AP_L)
YOLO	normal	normal	-	0.802	0.335	0.573	0.352	0.195	0.364	0.436
		low	MBLLEN	0.712	0.239	0.411	0.243	0.115	0.258	0.342
			KIND	0.729	0.254	0.437	0.255	0.138	0.293	0.365
			Zero-DCE	0.717	0.250	0.422	0.243	0.129	0.302	0.358
	low		-	0.764	0.318	0.522	0.309	0.162	0.344	0.405
MAET (w/o ort)	low+normal			0.770	0.321	0.534	0.331	0.163	0.355	0.401
MAET (w ort)				**0.788**	**0.330**	**0.569**	**0.341**	**0.189**	**0.362**	**0.421**

Table 2. The experimental results on VOC [6] dataset and COCO [23] dataset.

4. Experiments

4.1. Training Details

We realize our work based on the open-source object detection toolbox MMDetection [5]. The loss weight components, ω_1 and ω_2, in Eq. 5 are set to 1 and 10, respectively. In this experiment, L_{obj} represents the loss function of YOLO Head output branch in D_{obj}, L_{deg} represents the MSE loss of transformation parameters between the prediction of D_{deg} and the known ground truth, as listed in the last line of Table 1: (k, $\frac{1}{B}$, $\frac{1}{g_r}$, $\frac{1}{g_b}$, $\frac{1}{\gamma}$), each parameter is normalized in its corresponding category as a pre-process step, and their weights in L_{deg} are set to $5:1:1:1:1$.

All the input images have been cropped and resized to 608 × 608 pixel size. The backbone, DarkNet-53, is initialized with an ImageNet pre-trained model. We adopt stochastic gradient descent (SGD) as an optimizer and set the image batch size to 8. We set the weight decay to 5e-4 and momentum to 0.9. The learning rate of the encoder (E) and object detection decoder (D_{obj}) is initially set to 5e-4, and the learning rate of degrading transformation decoder (D_{deg}) is initially set to 5e-5. Both the rates adopt a MultiStepLR policy for learning rate decay. For the VOC dataset, we trained our network with a single Nvidia GeForce RTX 3090 GPU for 50 epochs, and the learning rate decreased by 10 at 20 and 40 epochs, and for the COCO dataset, we trained our network with four Nvidia GeForce RTX 3090 GPUs for 273 epochs, and the learning rate decreased by 10 at 218 and 246 epochs.

4.2. Synthetic Evaluation

Pascal VOC [6] is a well-known dataset with 20 categories. We train our model based on the VOC 2007 and VOC 2012 train and validation sets, and test the model based on the VOC 2007 test set. For VOC evaluation, we report mean average precision (mAP) rate at IOU threshold of 0.5.

COCO [23] is another widely used dataset with 80 categories and over 10,0000 images. We train our model based on the COCO 2017 train set and test the model based on the COCO 2017 validation set. For COCO evaluation, we evaluated each index of COCO dataset. The quantitative results of VOC and COCO datasets are listed in Table 2.

In this part, we train and test the YOLOv3 model [39] based on the VOC and COCO datasets for normal-lit and synthetic low-lit images as reference. Then, we use the YOLOv3 model trained for normal-lit to test on the set recovered by different low-light enhancement methods [28, 51, 8][2]. To verify the effectiveness of the orthogo-

[2]Here [28, 51] have been retrained on the normal-lit image and syn-

	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	**Total**
YOLO (N)	0.718	0.645	0.639	0.816	0.768	0.554	0.497	0.568	0.638	0.618	0.657	0.405	0.627
MBLLEN [28] + YOLO (N)	0.732	0.644	0.672	0.892	0.770	0.607	0.571	0.661	0.697	0.634	0.697	0.439	0.668
KIND [51] + YOLO (N)	0.734	0.681	0.655	0.862	0.783	0.630	0.569	0.627	0.682	0.671	0.696	0.482	0.673
Zero-DCE [8] + YOLO (N)	0.795	0.713	0.704	0.890	0.807	0.684	**0.657**	0.686	**0.754**	0.672	0.762	0.511	0.720
YOLO (L)	0.782	0.708	0.723	0.881	0.807	0.679	0.624	0.705	0.748	0.694	0.758	0.509	0.716
MAET (w/o ort)	0.792	0.711	0.730	0.884	0.811	0.671	0.648	0.701	0.750	0.702	0.754	0.514	0.722
MAET (w ort)	**0.813**	**0.716**	**0.745**	**0.897**	**0.821**	**0.695**	0.655	**0.726**	**0.754**	**0.727**	**0.774**	**0.533**	**0.740**

Table 3. Experimental results based on ExDark [25] dataset. YOLO (N), YOLO (L) are the models pretrained using original images/synthetic low-light images and fine-tuned based on the ExDark dataset; MBLLEN [28], KIND [51], and Zero-DCE [8] + YOLO (N) are pre-trained using the original COCO dataset and fine-tuned based on the Exdark dataset processed by different enhancement methods; MAET is our MAET (COCO) finetuned on the Exdark dataset.

nal loss, we train the MAET model with/without orthogonal loss function as MAET (w/o ort) and MAET (w ort) and directly test these models on the low-lit images with no pre-processing. To ensure fairness, all the methods in the training process are set to same setting parameters, *i.e.*, the data augmentation methods (expand, random crop, multisize, and random flip), input size, learning rate, learning strategy, and training epochs. Experimental configurations and results are listed in Table 2.

The experimental results in Table 2 show that our MAET has significantly improved the baseline detection framework based on the synthetic low-light dataset. Compared with the enhancement methods [28, 51, 8], the proposed MAET shows superior performance considering all evaluation indexes.

4.3. Real-World Evaluation

To evaluate the performance in a real-world scenario, we have evaluated our trained model (explained in Sec. 4.2) using the exclusively dark (ExDark) dataset [25]. The dataset includes 7,363 low-light images, ranging from extremely dark environments to twilight with 12 object categories. The local object-bounding boxes are annotated for each image. As EXDark is divided based on different categories, 80% samples of each category are used for fine-tuning on COCO pre-trained model (Sec.4.2) for 25 epochs with a learning rate of 0.001, and the remaining 20% are used for evaluation; we calculate the average precision (AP) of each category (see Table. 3 for more details) and calculate the overall mean average precision (mAP). Moreover, we have provided some examples in Appendix.A. As listed in Table 3, we can see that the proposed MAET method achieves satisfactory performance considering most of the classes and overall mAP. This result affirms that our degrading transformation is in accordance with real-world conditions.

Furthermore, we have evaluated our methods with the UG^2+ DARK FACE dataset [48]; UG^2+ is a low-light face detection dataset, which contains 6,000 labeled low-light face images, where 5400 images are used for fine-tuning on the COCO pretrained model (Sec.4.2) for 20 epochs with a learning rate of 0.001. The other 600 images are used for evaluation; the experiment results are listed in Table 4. The proposed MAET method has achieved better results compared with the other methods.

	mAP
YOLO (N)	0.483
MBLLEN [28] + YOLO(N)	0.516
KIND [51] + YOLO(N)	0.516
Zero-DCE [8] + YOLO(N)	0.542
YOLO (L)	0.540
MAET (w/o ort)	0.542
MAET (w ort)	**0.558**

Table 4. The experiment results on the UG^2+ DARK FACE [48] dataset.

5. Conclusion

We propose MAET, a novel framework to explore the intrinsic representation that is equivariant to degradation caused by changes in illumination. The MAET decodes this self-supervised representation to detect objects in a dark environment. To avoid the over-entanglement of object and degrading features, our method develops a parametric manifold along which multitask predictions can be geometrically formulated by maximizing the orthogonality among the tangents along the output of respective tasks. Throughout the experiment, the proposed algorithm outperforms the state-of-the-art models pertaining to real-world and synthetic dark image datasets.

Acknowledgement

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 and JST, ACT-X Grant Number JPMJAX190D, Japan and the National Natural Science Foundation of China under grant 62071333, U1830103, CSTC2018JSCX-MSYBX0115, China

thetic low-lit image pairs and [8] only trained on low-lit images, because [8] is a self-supervised model which do not need normal-lit ground truth.

References

[1] Chapter 5 - comparison of color demosaicing methods. In Peter W. Hawkes, editor, *Advances in Imaging and electron Physics*, volume 162 of *Advances in Imaging and Electron Physics*, pages 173–265. Elsevier, 2010.

[2] Josue Anaya and Adrian Barbu. Renoir – a dataset for real low-light image noise reduction. *Journal of Visual Communication and Image Representation*, 51:144 – 154, 2018.

[3] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019.

[4] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to see in the dark. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018.

[5] Kai Chen, Jiaqi Wang, and et.al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[6] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010.

[7] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian. Practical Poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008.

[8] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1777–1786, 2020.

[9] X. Guo, Y. Li, and H. Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017.

[10] Felix Heide and Steinberger et.al. FlexISP: A flexible camera image processing framework. *ACM Trans. Graph.*, 33(6), Nov. 2014.

[11] H. Jiang, Q. Tian, J. Farrell, and B. A. Wandell. Learning the image processing pipeline. *IEEE Transactions on Image Processing*, 26(10):5032–5042, 2017.

[12] 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.

[13] D. J. Jobson, Z. Rahman, and G. A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, 6(7):965–976, 1997.

[14] Hakki Can Karaimer and Michael S. Brown. A software platform for manipulating the camera imaging pipeline. In *European Conference on Computer Vision (ECCV)*, 2016.

[15] G. Kim, D. Kwon, and J. Kwon. Low-lightGAN: Low-light enhancement via advanced generative adversarial network with task-driven training. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2811–2815, 2019.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[17] Roman. Kvyetnyy, Roman. Maslii, and et.al. Object detection in images with low light condition. In Ryszard S. Romaniuk and Maciej Linczuk, editors, *Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2017*, volume 10445, pages 250 – 259. International Society for Optics and Photonics, SPIE, 2017.

[18] Edwin H. Land. An alternative technique for the computation of the designator in the retinex theory of color vision. *Proceedings of the National Academy of Sciences of the United States of America*, 83(10):3078–3080, 1986.

[19] C. Lee, C. Lee, and C. Kim. Contrast enhancement based on layered difference representation of 2D histograms. *IEEE Transactions on Image Processing*, 22(12):5372–5384, 2013.

[20] Chongyi Li, Jichang Guo, Fatih Porikli, and Yanwei Pang. Lightennet: A convolutional neural network for weakly illuminated image enhancement. *Pattern Recognition Letters*, 104:15–22, 2018.

[21] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T. Barron, Dillon Sharlet, Ryan Geiss, Samuel W. Hasinoff, Yael Pritch, and Marc Levoy. Handheld mobile photography in very low light. *ACM Trans. Graph.*, 38(6), Nov. 2019.

[22] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[23] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[24] Yajing Liu, Xinmei Tian, Ya Li, Zhiwei Xiong, and Feng Wu. Compact feature learning for multi-domain image classification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7186–7194, 2019.

[25] 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.

[26] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650 – 662, 2017.

[27] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *International Journal of Computer Vision*, 129(7):2175–2193, 2021.

[28] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. MBLLEN: Low-light image/video enhancement using CNNs. In *British Machine Vision Conference (BMVC)*, 2018.

[29] Yasushi Makihara, Masao Takizawa, Yoshiaki Shirai, and Nobutaka Shimada. Object recognition under various lighting conditions. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, pages 899–906, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[30] H. Nada, V. A. Sindagi, H. Zhang, and V. M. Patel. Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10, 2018.

[31] Lukas Neumann, Michelle Karg, Shanshan Zhang, Christian Scharfenberger, Eric Piegert, Sarah Mistr, Olga Prokofyeva, Robert Thiel, Andrea Vedaldi, Andrew Zisserman, and Bernt Schiele. NightOwls: A pedestrians at night dataset. In *Asian Conference on Computer Vision*, 2018.

[32] J. Nishimura, T. Gerasimow, R. Sushma, A. Sutic, C. Wu, and G. Michael. Automatic ISP image quality tuning using nonlinear optimization. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2471–2475, 2018.

[33] 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.

[34] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[35] T. Plotz and S. Roth. Benchmarking denoising algorithms with real photographs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2750–2759, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[36] C. Poynton, Inc Books24x7, and Engineering Information Inc. *Digital Video and HD: Algorithms and Interfaces*. Computer Graphics. Elsevier Science, 2003.

[37] Guo-Jun Qi, Liheng Zhang, Chang Wen Chen, and Qi Tian. AVT: Unsupervised learning of transformation equivariant representations by autoencoding variational transformations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8130–8139, 2019.

[38] R. Ramanath, W. E. Snyder, Y. Yoo, and M. S. Drew. Color image processing pipeline. *IEEE Signal Processing Magazine*, 22(1):34–43, 2005.

[39] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 91–99, Cambridge, MA, USA, 2015. MIT Press.

[41] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark - domain adaptation method for merging multiple models -. In *Proceedings - European Conference on Computer Vision*, August 2020.

[42] Liang Shen, Zihan Yue, Fan Feng, Quan Chen, Shihao Liu, and Jie Ma. Msr-net:low-light image enhancement using deep convolutional network, 2017.

[43] Mihai Suteu and Yike Guo. Regularizing deep multi-task networks using orthogonal gradients. *arXiv preprint arXiv:1912.06844*, 2019.

[44] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 839–846, 1998.

[45] Xiao Wang, Daisuke Kihara, Jiebo Luo, and Guo-Jun Qi. EnAET: Self-trained ensemble autoencoding transformations for semi-supervised learning. *arXiv preprint arXiv:1911.09265*, 2019.

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

[47] Xin Xu, Shiqin Wang, Zheng Wang, Xiaolong Zhang, and Ruimin Hu. Exploring image enhancement for salient object detection in low light images. *arXiv preprint arXiv:2007.16124*, 2020.

[48] W. Yang, Y. Yuan, and et.al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.

[49] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020.

[50] 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 Conference on Computer Vision and Pattern Recognition*, pages 2547–2555, 2019.

[51] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. MM '19, New York, NY, USA, 2019. Association for Computing Machinery.

[52] Y. Zheng, M. Zhang, and F. Lu. Optical flow in the dark. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6748–6756, 2020.

[53] Ziqiang Zheng, Yang Wu, Xinran Han, and Jianbo Shi. ForkGAN: Seeing into the rainy night. In *The IEEE European Conference on Computer Vision (ECCV)*, August 2020.

[54] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019.