Contents lists available at ScienceDirect

journal homepage: www.elsevier.com/locate/...

ELSEVIER

Comparing Deep Learning Models for Low-light Natural Scene Image Enhancement and their Impact on Object Detection and Classification: Overview, Empirical Evaluation, and Challenges

Rayan Al Sobbahi[a], Joe Tekli[a,*]

[a] *Department of Electrical and Computer Engineering,*
Lebanese American University (LAU), 36 Byblos, Mount Lebanon, Lebanon

Abstract

Low-light image (LLI) enhancement is an important image processing task that aims at improving the illumination of images taken under low-light conditions. Recently, a remarkable progress has been made in utilizing deep learning (DL) approaches for LLI enhancement. This paper provides a concise and comprehensive review and comparative study of the most recent DL models used for LLI enhancement. To our knowledge, this is the first comparative study dedicated to DL-based models for LLI enhancement. We address LLI enhancement in two ways: i) standalone, as a separate task, and ii) end-to-end, as a pre-processing stage embedded within another high-level computer vision task, namely object detection and classification. The paper consists of six logical parts. First, we provide an overview of the background and literature in LLI enhancement. Second, we describe the test data and experimental setup of the study. Third, we present a quantitative and qualitative comparison of the visual and perceptual quality achieved by 10 of the most recent DL-based LLI enhancement models. Fourth, we present a comparative analysis for object detection and classification performance achieved by 4 different object detection models applied on LLIs and their enhanced counterparts. Fifth, we perform a feature analysis of DL feature maps extracted from normal, low-light, and enhanced images, and perform the occlusion experiment to better understand the effect of LLI enhancement on the object detection and classification task. Finally, we provide our conclusions and highlight future steps and potential directions.

Keywords: Image Enhancement ; Low-light Conditions ; Deep Learning Models ; Object Detection and Classification ; Empirical Comparison

1. Introduction

With the rapid spread of digital devices and photo-taking gadgets such as smart phones, pads, and tablets, capturing digital images has become an easy and common trend in our world. These images may be influenced by various poor

* Corresponding author. Tel.: +961-9-547-262; fax: +961-9-546-262; email: joe.tekli@lau.edu.lb. Joe Tekli is also an adjunct researcher with the SPIDER research team, LIUPPA Laboratory, University of Pay and Pays Adour (UPPA), 64600, Anglet, Aquitaine, France.

Please cite this article as: First author et al., Article title
http://dx.doi.org/10.1016/j.....2017.00.000

visibility conditions like low-light, noise, haze, snow, and blur, among others. Low-light is a prominent element of our daily life that largely impacts the effectiveness of our vision and our ability to perceive meaningful content from badly illuminated objects, especially from Low-Light Images (LLIs). Low-light conditions are affected by the time of day (e.g., nighttime or twilight), the location (e.g., indoor or outdoor), and the availability of adequate light sources (e.g., natural and man-made lights) (Loh and Chan, 2019). Hence, LLI enhancement has emerged: i) as a standalone image processing task that aims at illuminating LLIs and improving their visual quality, and ii) as a pre-processing step embedded within another high-level computer vision task – like object detection and classification – to improve its performance.

Numerous traditional enhancement techniques have been proposed to tackle the LLI enhancement task. Histogram equalization methods like CLAHE (Pisano et al., 1998) and DHE (Abdullah-Al-Wadud et al., 2007) stretch the histogram of the image to make it uniform and enhance its contrast. Yet, they ignore spatially varying lightness and usually result in under or over brightened regions. Other methods rely on Retinex theory (Land and McCann, 1971) which splits the image into reflectance and illumination components: reflectance describes the intrinsic properties of the image's objects and is assumed to be constant under varying light conditions, while illumination represents the varying lightness in the whole image. Approaches like LIME (Li et al., 2015) and SRIE (Fu et al., 2016) adopt Retinex theory to perform LLI enhancement defined as an illumination estimation problem. Yet most Retinex-based approaches assume that enhancement does not affect image reflectance, regardless of the color distortions or lost details that result from applying the Retinex model (Wang et al., 2019). Also, Retinex-based enhancement quality is highly dependent on a set of hand-crafted parameters allowing to estimate the resulting illumination map (Wei et al., 2018).

Deep Learning (DL) approaches have been recently utilized to enhance LLIs and have shown great success. These approaches are data driven as they require training datasets of LLIs and their corresponding Normal-Light Images (NLIs). Yet most DL approaches face two major challenges (Yang et al., 2020): i) the data aspect challenge – state of art enhancement models mainly rely on synthetic training datasets which might not be well representative of real world LLIs that incorporate nonlinear and complex degradations due to their visual quality; and ii) the goal aspect challenge - LLI enhancement is usually embedded as a pre-processing step in another high-level computer vision task, while the enhancement model itself is not initially designed for the target task. One major question is whether a LLI enhancement method – which performs well as a standalone component – can improve (or not) the performance of the high-level computer vision task as a whole.

In this paper, we provide a concise and comprehensive review and comparative study of the most recent DL models used for LLI enhancement. We first briefly describe and categorize the different techniques and methods related to the problem, while illustrating some of their main characteristics. Then, we empirically compare the models in two ways: i) standalone, as a separate task by analyzing the visual and perceptual performance of 10 publicly available enhancement models, and ii) end-to-end, as a pre-processing stage embedded within another high-level computer vision task: by comparing the performance of 4 object detection and classification models, applied on images enhanced by each of the 10 LLI enhancement models considered in the previous experiment. We also perform a DL feature analysis experiment to compare the feature maps extracted from LLIs, NLIs and enhanced images, and run the occlusion experiment (Zeiler and Fergus, 2014) to better understand the effect of LLI enhancement on preserving the semantic features needed by the object detection and classification task. To our knowledge, this is the first comparative study dedicated to DL-based models for LLI enhancement, and we hope the obtained results will foster and guide further research on the subject.

The remainder of this paper is organized as follows: Section 2 provides an overview of LLI enhancement techniques. Section 3 describes and categorizes some of the most prominent DL-based LLI enhancement models. Section 4 describes the test data and experimental setup. Section 5 presents the results of Experiment 1: comparing visual and perceptual LLI enhancement quality, Section 6 covers Experiment 2: comparing object classification and detection quality, and Section 7 covers Experiment 3: comparing image feature maps and occlusion results. Section 8 presents a recap of the results and highlights interesting future directions, before concluding in Section 9.

2. Overview of LLI Enhancement

The main objective of LLI enhancement is to improve the visual quality of an image by boosting its illumination and contrast while avoiding amplified noise or exposed artifacts. Formally, a low-light image I_{LLI} is the output of a degradation function:

$$I_{LLI} = D\ (I_{NLI}, \delta) \tag{1}$$

where D denotes a degradation mapping function, I_{NLI} the NLI, and δ the parameter of the degradation process (e.g. illumination level). Generally, the degradation process is complex as it may encompass – in addition to the illumination of the image – other factors like artifacts and noise. The enhancement task aims at recovering an approximation of I_{NLI} denoted by $I_{Enhanced}$, generated from I_{LLI} as follows:

$$I_{Enhanced} = F\ (I_{LLI}, \theta) \tag{2}$$

where F is the LLI enhancement model and θ its adjustment parameters. Here, we distinguish between two main categories of LLI enhancement models: i) traditional and ii) deep learning.

2.1. Traditional Approaches

Most traditional LLI enhancement techniques rely on mathematical or algorithmic models to perform the enhancement task. They fall under four main categories: i) gamma correction, ii) histogram equalization, iii) Retinex theory based, and iv) frequency-based methods. Gamma correction methods, e.g., (Huang et al., 2013; Zhi et al., 2018) use a nonlinear transformation-based function in which a gamma correction parameter is adjusted to stretch or compress different gray regions of the image, aiming to enhance it. Histogram equalization methods, e.g., (DHE by Abdullah-Al-Wadud et al. (2007); CLAHE by Pisano et al. (1998); Wang et al., 1999) rely on a cumulative distribution function to change the image output gray levels such that they fit into a uniform distribution. The original LLI is mapped to its enhanced counterpart with an approximately uniform gray-level distribution. Yet, gamma correction and histogram equalization methods generally ignore spatially varying lightness and usually result in under or over brightened regions.

Retinex theory, i.e., the theory of the human retinal cortex (Land and McCann, 1971), has been utilized to perform LLI enhancement. Based on the nature of color perception by the human eye and the modeling of color constancy, methods in this category aim to remove the effects of illumination from the image leaving it with the reflective nature of its objects (Jobson et al., 1997a; Rahman et al., 1996; Jobson et al., 1997b, Hao et al., 2020). According to the theory, the Human Visual System (HVS) perceives the content and colors of the image constantly under varying or uneven lighting conditions, and thus only the major characteristics of the objects depicted in the reflection component are retained by the HVS (Lee et al., 2015). As a result, the reflectance component of the image is considered to be constant under varying light conditions and holds the inherent characteristics of visual objects. The Retinex model is thus used to estimate the illumination component of the image and retain its reflectance component, preserving the image's inherent features to allow more accurate image processing. The authors in (Zosso et al., 2015) introduce a generalized Retinex model by reinterpreting the gradient thresholding model as variational models with sparsity constraints. The authors define a filtered gradient as the solution of an optimization problem considering sparsity and fidelity prior of the reflectance gradient to the observed image gradient. Consequently, they fit the actual reflectance gradient to it, and adapt it to further sparsity and fidelity priors. This generalized model allows making connections with other variational or kernel-based Retinex implementations. Yet most Retinex-based approaches like MultiScale Retinex (MSR) (Rahman et al., 1996), MultiScale Retinex with Color Restoration (MSRCR) (Jobson et al., 1997b), Single Scale Retinex (SSR) (Jobson et al., 1997a), LIME (Li et al., 2015) and SRIE (Fu et al., 2016) assume that enhancement does not affect image reflectance, regardless of the color distortions or lost details that result from applying the Retinex model (Wang et al., 2019). In addition, Retinex-based enhancement quality is highly dependent on a set of carefully hand-crafted parameters allowing to estimate the resulting illumination map (Wei et al., 2018). A recent approach proposed in (Khan et al., 2021a) aims at mitigating some of the limitations of Retinex-based techniques. It detects and matches feature points across adjacent viewpoints, and determines the exposure gain among the matched feature points. This is used to design an image restoration method to restore multi-view low dynamic range images for each viewpoint, and then fuses them together to produce high-quality images for each viewpoint without capturing a series of bracketed exposure.

Frequency domain-based methods have also been utilized to perform LLI enhancement, e.g., (Xiao et al., 2008; Zhang et al., 2013). They transform images from their initial spatial domain to a frequency domain using Fourier analysis. Filtering is then applied on the frequency-based representations to perform enhancement, before the final images are brought back to their initial spatial domain. More recent techniques rely on Homographic Filtering (HF) to perform frequency-based enhancement, e.g., (Zaheeruddin and Suganthi, 2019; Zhang et al., 2018; Han et al, 2009). In HF, the illumination and reflectance components of the image are converted into the frequency domain in which a high-pass filter is used to enhance high-frequency reflection components and suppress low-frequency illumination components. In (Zhang et al, 2018), the authors integrate a denoising approach based on guided image HF to eliminate the amplified noise and avoid the edge blocking effect. The authors in (Han et al, 2009) design a two-channel HF image enhancement method, where they first convert the input image from RGB to HSV color domain, and then perform enhancement separately on the saturation channel (S) and on the illumination channel (V) using Butterworth HF and Gaussian HF respectively. In (Zaheeruddin and Suganthi, 2019), the authors combine HF with a parametric fuzzy transform. They first use a HF algorithm to acquire the illumination image of the V channel in the HSV domain, and then perform fuzzy image processing through a parametric transform to smooth and enhance the image's illumination. Yet, most HF-based methods require two (or more) Fourier transforms which are computationally expensive, thus limiting real-time image enhancement.

2.2. Deep Learning Approaches

In contrast to the traditional algorithmic or mathematical enhancement approaches, Deep Learning (DL) enhancement models are essentially data-driven, where training datasets of LLIs and NLIs are used to drive the learning process. DL models are a special kind of machine learning algorithms made of multilayered artificial neural networks, inspired by the structure and function of the human brain. They aim to find unknown structures or patterns in the input distribution so that they discover good representations of the data and learn its features through a hierarchical architecture (Deng, 2014). DL techniques have gained great attention in the past few years as the most effective machine learning solutions to perform LLI enhancement, outperforming traditional methods based on histogram equalization e.g. (Pisano et al., 1998; Abdullah-Al-Wadud et al., 2007); and Retinex theory e.g., (Jobson et al., 1997a; Jobson et al., 1997b; Rahman et al., 1996). They accept LLIs as input, and propagate them through the DL model to learn a variety of features needed for the enhancement task. Paired labels of LLIs/NLIs are essentially needed to train the DL model under a supervised setting, allowing it to learn how to perform the enhancement task. A loss function is one of the main elements of a DL solution, allowing to evaluate how well a given model fits the training data. Through an iterative self-evaluation process, the loss function usually guides the DL model to reduce the error in its own predictions. In this context, commonly used DL loss functions like Mean Absolute Error (MAE, or L1 loss) and Mean Square Error (MSE, or L2 loss) might not always be suitable to accurately evaluate the visual quality of enhanced LLIs (Wang et al., 2004). Given the various elements that affect the quality of the image including illumination levels, color deviations, artifacts, noise, etc., recent studies have introduced more sophisticated loss functions to improve the quality of enhanced LLIs, including: perceptual loss (Lv et al., 2018), illumination smoothness loss (Wei et al., 2018), and adversarial loss (Wang et al., 2019), among others.

DL techniques represent the current trend for research in the LLI enhancement task. Early DL solutions like LLNet (Lore et al., 2017) and LLCNN (Tao et al., 2017) have shown improved enhancement results compared with traditional approaches. Table 1 shows quantitative comparison results of LLCNN with different histogram equalization and Retinex-based methods. Results show that LLCNN takes the lead considering most assessment metrics. A qualitative comparison in Fig. 1 shows sample enhanced images produced by LLCNN and its traditional counterparts. LLCNN seems to produce a more natural-enhanced image with vivid colors and without apparent artifacts or over exposure.

While they usually require expensive training time and effort (Abu-Khzam et al., 2019; Abu-Khzam et al., 2015), yet various reasons have contributed to the leap of DL algorithms and their applications, including (Deng, 2014): i) the substantial increase in computational capabilities (e.g., GPUs), ii) the lower costs of computing hardware, iii) the significant advances of machine learning algorithms (Salem et al., 2018; Ebrahimi et al., 2021; Abu-Khzam et al., 2018), and iv) the increasing availability of training data. In the following section, we describe and categorize the recent DL models for LLI enhancement.

Table 1. Quantitative comparison of traditional and early DL-based methods on the LLCNN datasets (Tao et al., 2017) (red color refers to the best score and green to the second best score for every metric)

Approach	***PSNR*** ↑	***SSIM*** ↑	***LOE*** ↓	***SNM*** ↑
LLCNN	**35.20**	**0.957**	**10.27**	**0.598**
CLAHE	16.54	0.686	75.73	0.346
DHE	13.35	0.546	58.30	0.195
LIME	19.17	0.719	92.70	0.565
SRIE	14.83	0.605	56.82	0.200
MSRCR	**25.97**	0.894	65.93	**0.612**
SSR	22.33	**0.908**	**25.93**	0.571
Dark	12.70	0.476	50.82	0.118

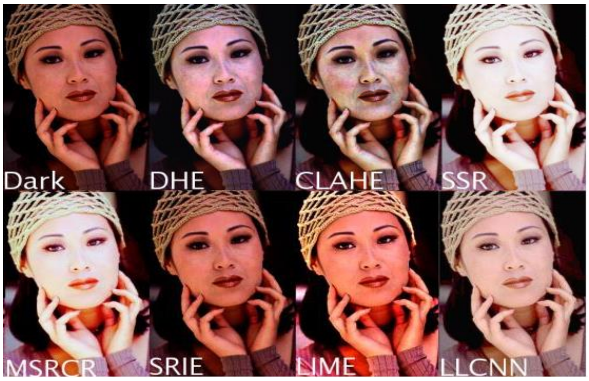


Fig. 1. Visual comparison of traditional and early DL-based methods (Tao et al., 2017)

3. Deep Learning-based LLI Enhancement

We organize DL-based LLI enhancement solutions in five main categories: i) Encoder-decoder and Convolutional Neural Network (CNN) based models, ii) Retinex theory-based models, iii) Fusion based models, iv) Generative Adversarial Network (GAN) based models, and more recent v) Zero Reference models.

3.1. Encoder-decoder and CNN-based Models

Various works have focused on utilizing encoder-decoder models, CNNs, or have integrated them together to perform LLI enhancement.

Encoder-decoder models: An encoder-decoder is a DL model designed to learn a mapping from an input domain to an output domain through a two-stage network comprising: i) an encoder which encodes the input into a latent feature representation, and ii) a decoder which decodes and reconstructs the original features to predict the output. While largely used in image-to-image translation applications (Minaee et al., 2020), encoder-decoder solutions have been recently developed for image enhancement, where the input is a LLI, and the output is its enhanced counterpart, e.g., (Lore et al., 2017; Jiang et al., 2018; Xu et al., 2018). An autoencoder is a special type of encoder-decoder which aims at learning a reduced encoding for the data, and to generate from the reduced encoding a representation as close as possible to its original input (Minaee et al., 2020). There are many variants of autoencoders such as: i) sparse autoencoders: extracting sparse features from the input data, by penalizing hidden unit biases (Ranzato et al., 2006) or unit activations (Le et al., 2011), ii) denoising autoencoders: recovering the correct input from a corrupted version of the input data, by forcing the network to learn the structure of the input distribution (Vincent et al., 2008), and iii) convolutional autoencoders: combining CNNs and autoencoders, where the encoder consists of a series of convolutional and pooling layers and the decoder consists of deconvolutional and unpooling layers.

LLNet (Lore et al., 2017) is one of the earliest DL approaches for LLI enhancement. It uses a stacked-sparse denoising autoencoder (SSDA) as its deep neural network architecture with three denoising autoencoder layers comprising hidden units with no use of convolutional layers. The model is trained on synthetic LLIs obtained from normal images through gamma correction and Gaussian noise induction, and uses the L2 loss function. Experimental results by Lore et al. (2017) highlight a tradeoff between the sharpness of the enhanced image and its noise levels. The model shows competitive results when compared with traditional approaches based on histogram equalization (Abdullah-Al-Wadud et al., 2007; Pisano et al., 1998) and gamma adjustment.

CNN models: A Convolutional Neural Network (CNN) is a DL network consisting of a regularized version of the multilayer perceptron that uses the linear convolution mathematical operation in place of general matrix multiplication in at least one of its layers. CNNs are highly effective and have been commonly used in various computer vision applications (Guo et al., 2016), allowing to extract, distinguish, and assemble complex visual features (patterns) from the images' visual properties and objects. A typical CNN consists of three types of consecutive layers: i) convolutional layers: using kernels to convolve the whole image as well as intermediate feature maps and generate new feature maps, ii) pooling layers: reducing the feature map dimensions and the number of network parameters, and iii) fully connected layers: mapping a 2D feature map into a 1D feature vector that either refers to a certain number of categories for image classification or is utilized for further processing. CNNs have been largely used for the image classification task, including famous architectures such as VGG16 and VGG19 (Simonyan and Zisserman, 2015), AlexNet (Krizhevsky et al., 2012), and ResNet (He et al., 2016).

LLCNN (Tao et al., 2017) is one of the early CNN-based models for LLI enhancement. It is built using specially designed convolutional modules inspired from inception modules (convolving an input using different size convolutional layers and then combining their outputs to the next layer) and residual modules (employing shortcut connections). It uses a Structural Similarity Index (SSIM) (Wang et al., 2004) based loss function and relies on synthetic LLIs created through random gamma adjustment for training the network. The model demonstrates superior performance compared with LLNet (Lore et al., 2017) and many traditional approaches (Abdullah-Al-Wadud et al., 2007; Pisano et al., 1998; Rahman et al., 1996; Jobson et al., 1997a; Fu al et., 2016).

Gharbi et al. (2017) propose a deep bilateral CNN based model to perform fast real-time enhancement. The approach aims at processing a low-resolution version of the image in which a bilateral grid of affine coefficients is estimated. Then a slicing operation is used to up-sample the affine coefficients into the full image resolution. The model is designed to learn global and local features and preserve edges. L2 loss is used to train the network on the MIT FiveK dataset (Bychkovsk et al., 2011). The results demonstrate the effectiveness of the model in real time image enhancement. One limitation mentioned by the authors is the network's strong dependence on the modeling assumptions and constraints related to the affine transformations in the bilateral space.

Chen et al. (2018) introduce a learning to *See In the Dark* (SID) model for image enhancement and noise suppression designed to process images under extreme low-light conditions. The model relies on Fully Convolutional Networks (FCNs) (using convolutional layers only) and is trained using L1 loss on a newly collected dataset of raw LLIs taken by the imaging sensors of Sony 7SII and Fujifilm X-T2 cameras. Although the model is able to suppress noise and produce proper coloring, it is limited to raw data obtained using a specific camera sensor and the images of the SID dataset do not contain pictures of humans and dynamic objects (Chen et al., 2018).

Integrated models: Jiang et al. (2018) propose LL-RefineNet, a deep refinement network consisting of two symmetrical paths: forward and backward. In the forward path, high-level features with global content are extracted and then gradually fused with low-level features with local content and refined during the backward refinement path. The model relies on synthetic LLIs based on impulse and Gaussian noise and guided using a mixed loss function of L1 and L2 losses. Results show that the model outperforms LLCNN (Tao et al., 2017) and many traditional approaches both quantitatively and qualitatively.

Xu et al. (2018) introduce LRCNN: a Low-light Residual Connection based Convolutional Network, consisting of: i) a convolutional encoder-decoder structure in which the encoder is used for feature extraction and the decoder for denoising, connected with a ii) sequence of fully connected layers for brightness enhancement. Residual connections are used to better preserve the details of the original image. The network is guided by an L2 based loss function and is trained on a synthetic dataset of LLIs simulated from the CVG-UGR database. Results show that the model can remove noise and properly adjust light intensity.

Wang et al. (2018) introduce a Global Illumination Aware and Detail-preserving NETwork (GLADNET) comprising: i) a global illumination estimation step using an encoder-decoder structure where the encoder consists of convolution layers and the decoder consists of resize convolutional layers (Odena et al., 2016), followed by ii) a reconstruction step through a series of convolutions where the input image is concatenated with the predicted features from the encoder-decoder to better preserve the original image features. The network is trained on a synthesized dataset collected from RAISE (Dang-Nguyen et al., 2015) and guided by L1 loss. Results show that the model produces clear and natural enhanced images with preserved details.

3.2. Retinex Theory-based Models

Other DL approaches are inspired by the Retinex theory (Land and McCann, 1971) that decomposes the image into a constant reflectance map and a light varying illumination map (cf. Section 2.1). Multi Scale Retinex Net (MSR-Net) (Shen et al., 2017) is one of the early models in this category. It performs LLI enhancement in three stages. The input LLI is first processed as a set of multi-scale logarithmic transformations. The transformed image is then fed into a CNN, and is finally processed through a dedicated color restoration function. The model is trained using the L2 loss function and a synthesized dataset obtained from the UCID dataset (Schaefer and Stich, 2003), the BSD dataset (Arbelaez et al., 2011), and Google Images. While the model is effective in producing images with rich colors and clear textures, yet it sometimes fails to properly handle the image edge features as it tends to produce some darkness around the edges, especially in bright regions (referred to as the “halo effect”) (Shen et al., 2017).

Another approach is RetinexNet (Wei et al., 2018) which consists of two subnetworks: i) DecomNet that aims at learning the decomposition of the image into its reflectance and illumination components based on Retinex theory, and ii) EnhanceNet that performs illumination adjustment and enhancement through a dedicated encoder-decoder structure which uses multiscale concatenation to maintain the global and local illumination of the enhanced image. A joint denoising operation using 3D transform-domain filtering (BM3D) denoising algorithm (Dabov et al., 2006) is then applied on the reflectance component. Wei et al. (2018) introduce their own training dataset named LOw-Light (LOL), consisting of 500 pairs of real LLIs and NLIs. They also put forth a multi-term loss function combining reconstruction, invariable reflectance, and illumination losses. The resulting enhanced images are produced with a good image decomposition learning and are deemed visually pleasing by the authors.

Li et al. (2018) introduce LightenNet, a CNN model made of 4 convolutional layers for i) patch extraction and representation, ii) feature enhancement, iii) non-linear mapping, and iv) reconstruction. It is designed to predict the Retinex illumination map component from the original LLI, which is then used to produce the enhanced image. The network learns through a synthesized dataset obtained by the Retinex model and is guided by the L2 loss function. The enhanced images are visually pleasing with well restored content. Yet, the method shows a degraded performance while applied on low-quality images due to noise or JPG compression resulting in noise and artifacts amplification (Li et al. 2018).

Zhang et al. (2019) propose KinD (Kindling the Darkness) consisting of three networks: i) layer decomposition that decomposes the image into reflectance and illumination components, ii) reflectance restoration which aims at removing degradations that are concentrated in the dark regions of the reflectance, and iii) illumination adjustment which distributes the illumination across the image. The authors design an integrated loss function based on L1, L2, and SSIM (Wang et al., 2004) losses, and train the model on the LOL dataset (Wei et al., 2018). Results in (Zhang et al., 2019) show that the model produces enhanced images with properly adjusted lightness and suppressed noise.

Wang et al. (2019) introduce a Deep Underexposed Photo Enhancement (DeepUPE) model which performs image-to-illumination map learning. It consists of an encoder network (i.e., a pre-trained VGG16 (Simonyan and Zisserman, 2015) that extracts the image’s local and global features, followed by a bilateral grid based up-sampling allowing to produce the image’s full resolution illumination map. The latter is then used to enhance the image based on the Retinex model. The authors use an integrated loss function combining reconstruction, smoothness, and color losses. A newly proposed dataset of underexposed images and expert retouched references is used for training and evaluation. Results by Wang et al. (2019) show a good recovery of the image details, contrast, and colors.

The authors in (Khan et al., 2021b) decompose images into reflection and illumination maps, which are respectively used to solve the illumination blindness and structure degradation problems. The hidden degradation in reflection and illumination are tuned with a knowledge-based adaptive enhancement constraint designed for ill-illuminated images. The model maintains a balance of smoothness and contributes to solving the problem of noises as well as over- and under-enhancement. The local consistency in illumination is achieved by a repairing operation performed using a dedicated Repair-Net model. The total variation operator is optimized to acquire local consistency, and the image gradient is guided with the enhancement constraint. The enhanced image is obtained as a product of the updated reflection and illumination maps. Results on a new dataset show improved results in preserving structural and textural details compared with existing solutions.

Another approach in (Khan et al., 2021c, Khan et al., 2021d) introduces FSDG-Net (Few-Shot Divide and Glow network) which aims to enhance low and ill-lighting images. The network comprises: i) Multilayer Image Division-Net (MID-Net) which splits the image into reflection and illumination transmission components based on the Retinex model, and ii) Glow-Net which boosts the illumination map of the image through an encoder-decoder model. A contrast enhancement strategy is adopted which allows training the network from the correlation consistency of the input image decomposition itself, thus relying on few training data samples. The network is driven by multiple losses including an optimized total variation loss and a structure and texture loss, among others. A newly designed dataset: D3I-dataset (consisting of 880 images) is used for training and testing. Results show a good preservation of colors and contrast and proper handling for uneven illumination in input images.

3.3. Fusion-based Models

Some LLI enhancement models consider fusing the derived images or feature maps by multiple traditional or DL techniques to combine their advantages into a final enhanced image. MBLLEN, a Multi-Branch Low-Light Enhancement Network (Lv et al., 2018) is one of the earliest models in this category. It uses a dedicated feature extraction module to extract the LLI features at each of its 10 convolutional layers, and then enhances the features at each layer using an encoder-decoder based enhancement module. It finally fuses the multi-branch enhanced features to form the enhanced image. The model uses a loss function composed of structure, context, and region losses. It learns through a synthesized dataset of LLIs from the Pascal VOC dataset (Everingham et al., 2012). The resulting enhanced images have good brightness and contrast with minimal artifacts.

Shin et al. (2018) propose ACA-net, an Adversarial Context Aggregation network consisting of a Context Aggregation Network (CAN) applied with an adversarial GAN-based loss function. First, image illumination is boosted using two gamma correction functions, then the corresponding feature maps are extracted using convolutional layers and passed through a CAN which uses dilated convolutions to perform an effective aggregation of the global contextual information in the image. The network is guided by an integrated loss function combining reconstruction and adversarial losses, and the training data is synthesized based on aesthetic visual analysis (AVA) dataset (Murray et al., 2012). The model shows superior performance compared with MSR-Net (Shen et al., 2017).

Another fusion-based approach is DFN (Deep Fusion Network) (Cheng et al., 2019), which combines three traditional enhancement techniques: CLAHE (Pisano et al., 1998), log correction and bright channel enhancement. It runs the three models on the same input LLI and produces the feature confidence maps from the three derived images using an encoder-decoder network. It then weights the derived images by the obtained confidence maps in an element-wise fusion to output the final enhanced image. The model aims at combining the significant features emphasized by the constituent enhancement methods. It utilizes an integrated loss function composed of L1 and L2 losses and is trained on a dataset synthesized from 600 NLIs using gamma correction. Although it shows good performance, yet the authors mention that DFN may add smoothing and artificial edges in the fused image as it lacks an edge preserving capability (Cheng et al., 2019).

Wang et al. (2019) utilize attention modules to selectively enhance useful features while suppressing features that are not so important for the network, and use multi-scale feature fusion to combine global features with strong semantic features at deeper layers. The model consists of feature extraction blocks (FEB), where each FEB is a convolutional block made up of an attention module and two convolutional layers, and a feature fusion block (FFB) which fuses multilevel features through pixel-wise addition and channel connection. The model uses a Peak Signal to Noise Ratio (PSNR) based loss function and is trained on real images from the SID (Chen et al., 2018) and S7ISP (Schwartz et al., 2019) datasets, as well as synthetic images produced based on the Pascal VOC dataset (Everingham et al., 2012). The model shows competitive results compared with many traditional enhancement approaches (Jobson et al., 1997b; Fu et al., 2016; Li et al., 2015).

Lv et al. (2020) introduce an attention-guided model that aims at handling image enhancement and denoising simultaneously by using Under Exposure (UE) attention maps and noise maps that guide the model attention in a region-aware adaptive manner. The model consists of four components: i) Attention Net: produces the UE maps used to avoid over-enhanced regions, ii) Noise Net: estimates the noise distribution map, iii) Enhancement Net: extracts and enhances features then fuses them through a multi-branch CNN concatenation and iv) Reinforce Net: uses dilated convolutions to improve the image contrast and details. The integrated loss function combines L1, L2, SSIM (Wang et al., 2004), and VGG19 (Simonyan and Zisserman, 2015) based perceptual losses, among others. The network is trained on a synthesized dataset from publicly available datasets like Pascal VOC (Everingham et al., 2012) and Microsoft (MS) COCO (Lin et al., 2014). Extensive evaluation experiments show the superior performance of the proposed model compared with LLNet (Lore et al., 2017), MBLLEN (Lv et al., 2018), SRIE (Fu et al., 2016), LIME (Li et al., 2015), among others.

Ren et al. (2019) propose a deep hybrid network consisting of two streams that simultaneously learn: i) the global content and ii) the salient edge contents of the input image. The first stream uses a residual encoder-decoder and the second stream utilizes a novel spatially variant recurrent neural network (RNN) to model the edge details. The network is guided by an integrated loss function combining L2, perceptual, and adversarial losses, and is trained using MIT-Adobe FiveK dataset (Bychkovsky et al., 2011). The enhanced images are shown to be visually pleasing with minimal artifacts and color distortions (Ren et al., 2019).

Xiang et al. (2019) introduce a multi-branch encoder-decoder architecture combining: i) DCGAP: a Dilated Convolution and Global Average Pooling module used to better learn the image global features, and ii) ConvLSTM: a Convolutional Long Short-Term Memory that allows remembering and preserving the features learned at the different branches. The model is guided by L1 and SSIM (Wang et al., 2004) losses and is trained using the LOL dataset (Wei et al., 2018) and 1000 synthetic images based on RAISE (Dang-Nguyen, 2015). The model successfully enhances LLI visual quality while minimizing noise and artifacts (Xiang et al., 2019).

3.4. GAN-based Models

Recently, Generative Adversarial Networks (GANs) have been attracting attention for image-to-image mapping applications (Goodfellow et al., 2014), and have been successfully employed for the LLI enhancement task. A typical GAN is made-up of two networks: a generator and a discriminator. The generative network is trained to generate realistic synthetic data samples from a data distribution of interest, while the discriminative network is trained to distinguish fake samples produced by the generator form the true data distribution. The generative network's training objective is to increase the error rate of the discriminative network, as it attempts to "fool" the discriminator network by producing novel candidates that the discriminator thinks are not synthesized. DL models like encoder-decoders and CNNs are used for the generator and discriminator networks. In the context of image enhancement, LLIs are used as real samples and enhanced images as fake samples to be generated.

Meng et al. (2019) introduce one of the earliest GAN-based models to perform LLI enhancement, consisting of an encoder-decoder based generator supplemented by a fusion network that combines features from the different layers of the encoder-decoder. Through adversarial learning, the discriminator is trained to differentiate a LLI from an enhanced image while the generator is trained to fool the discriminator. The model learns using a vehicle dataset of daytime and nighttime images that are not exactly taken at the same scenes, and is driven by an integrated loss function combining adversarial, perceptual, and total variation losses. A major problem highlighted by the authors is the tendency of the model to miss objects that are strongly illuminated in nighttime images.

Hua and Xia (2018) propose a GAN-based approach supported by Image Quality Assessment (IQA) techniques, in particular an image quality assessment network NIMA (Talebi and Milanfar 2018) which relies on the VGG16 (Simonyan and Zisserman, 2015) feature extractor to minimize the model’s dependence on the training dataset and boost its de-noising and de-blurring performance. The authors use an integrated loss function combining IQA, content, and total variation losses, and introduce a synthesized dataset based on General100 (Dong et al., 2016) and other image sources by applying Gaussian correction, Gaussian blur, and noise induction techniques on the normal images. Results in (Hua and Xia, 2018) highlight a certain balance between noise suppression and the preservation of image details.

Kim et al. (2019) introduce Low-LightGAN which applies spectral normalization on the network to make the training more stable and accurate. It uses a combination of loss functions including adversarial, perceptual, color, and total variation losses, specifically tuned to produce visually pleasing images. The authors propose a task-driven training dataset based on local illumination synthesis rather than global low-light synthesis, so that over-saturated bright regions in the image are avoided. Results show good performance although the model may add artifacts in the background of the enhanced images.

Yangming at al. (2019) combine Retinex theory and GANs. Their generative network includes: i) a decomposition part that decomposes the image into its reflectance and illumination components, and ii) an enhancement part that enhances the lightness of images taken from the CSID dataset (Chen et al., 2018). The loss function combines regularization, reconstruction, and adversarial losses, among others. Results by Yangming at al. (2019) show that combining Retinex theory and GANs can effectively handle LLI enhancement.

Wang et al. (2019) describe Retinex Decomposition based Generative Adversarial Network (RDGAN) which consists of two subnetworks: i) Retinex Decomposition Net (RDNet) that decomposes the LLI into its illumination and reflectance components, and ii) Fusion Enhancement Net (FENet) that fuses the decomposed parts into an enhanced image. The model is trained using the SICE dataset (Cai et al., 2018) and utilizes a novel adversarial loss function based on GANs to improve visual quality. While the model can properly recover the details and colors of the original LLI, yet it also tends to amplify noise and JPEG artifacts that are not obvious in the LLI, thus possibly degrading the quality of the enhanced image (Wang et al., 2019).

Chen et al. (2018) propose a Deep Photo Enhancer (DPE) model using a GAN-based architecture for image enhancement, while considering paired and unpaired training settings (i.e., with and without LLI/NLI pairs[1]). A global feature U-Net (Ronneberger et al., 2015) is used to investigate the paired training setting. Two network architectures are used for unpaired training: 1-way GAN and 2-way GAN. In addition, two improvements are added to stabilize the training: adaptive WGAN (Arjovsky et al., 2017) and individual batch normalization for the generator. The loss function is based on L2 and adversarial losses. The authors produce a dataset extracted from MIT-Adobe 5K (Bychkovsky et al., 2011) and HDR images selected from Flickr images. Results mainly show good quality enhanced images with natural colors, yet the authors also highlight that the model might amplify noise in very dark and noisy images.

A recent approach by Jiang et al. (2021) introduces EnlightenGAN: a first successful attempt at generalizing well to various real world scenes while using unsupervised learning for image enhancement based on GANs. The model undergoes unpaired training, uses an attention guided U-Net (Ronneberger et al., 2015) as the backbone for the generator, and includes a global relativistic discriminator (Jolicoeur-Martineau, 2018) along with a local one to handle spatially varying light conditions in the image. Self-regularization is adopted for the loss function and the attention mechanism, since the model is independent from reference training labels. The loss function combines local and global discriminator adversarial losses and a self-feature preserving loss. The training dataset consists of unpaired LLIs and NLIs sampled from the LOL (Wei et al., 2018), RAISE (Dang-Nguyen et al., 2015) and HDR datasets (Gharbi et al., 2017; Kalantari and Ramamoorthi, 2017). Results by Jiang et al. (2021) demonstrate a successful enhancement of dark areas while preserving the texture details and producing naturalistic images with no under- or over-exposed regions.

3.5. Zero Reference Models

A recent approach by Guo et al. (2020) opens the door for a new category of LLI enhancement techniques which does not require paired or unpaired training data (hence the name “Zero Reference”). The authors introduce Zero Reference Deep Curve Estimation (ZeroDCE) which entirely reformulates the LLI enhancement task: from an image-to-image mapping task into an image-to-light curves estimation task. Inspired by curve adjustment techniques used in digital photo editing solutions, the authors design light enhancement curves that are learned and estimated by a lightweight deep curve estimation network (DCE-Net), and are then iteratively applied on the input LLI to produce the final enhanced image. The model can be trained in the absence of paired or even unpaired training data by using non-reference loss functions such as spatial consistency, exposure control, color constancy, and illumination

[1] Unpaired training is increasingly used with GANs and consists in training the model using unmatched training data, e.g., LLIs and NLIs which are produced separately, and which do not necessary match.

smoothness losses that can indirectly evaluate the quality of enhancement. The proposed method is computationally efficient and shows superior performance compared with DL enhancement models like EnlightenGAN (Jiang et al., 2021), RetinexNet (Wei et al., 2018), and LIME (Li et al., 2015), among others.

Another approach by Zhang et al. (2020) presents a self-supervised DL model for LLI enhancement, which can be trained using only LLIs with no need for paired or unpaired training data. Relying on the Retinex model and entropy theory, the authors devise a well-tuned loss function that includes a new method to compute reflectance loss. The method is based on the assumption that the enhanced image should have enough information and should comply with the original image. This is achieved by applying histogram equalization on the LLI to improve its information entropy. Driven by the newly designed reference-less loss function, a CNN model is then trained on the LOL dataset (Wei et al., 2018) to perform the enhancement task. Results by Zhang et al. (2020) show that the model produces visually pleasing images with short training time, and exhibits good real-time performance.

A more recent approach by Ma et al. (2022) introduces a Self-Calibrated Illumination (SCI) learning framework that aims at predicting the illumination map of the original image through cascaded stages in which residual representations are learned. A self-calibrated module is designed to ensure the convergence of each stage to the same state, resulting in a progressive self-correction of the input image. The authors devise an unsupervised training loss composed of fidelity and smoothing losses. Extensive evaluations show that the model produces competitive results compared with state of the art solutions like ZeroDCE (Guo et al., 2020) and EnlightenGAN (Jiang et al., 2021) while enjoying the leanest architecture and the fastest inference time.

3.6. Discussion

Table 2 summarizes the main characteristics of recent DL-based LLI enhancement solutions. While many DL enhancement models have been shown to outperform their traditional counterparts, e.g., (Lore et al., 2017; Tao et al., 2017; Jiang et al., 2018), yet most of them share several challenges.

Challenge 1: Most approaches consider supervised learning where paired LLIs/NLIs are needed to train the models. Yet collecting large datasets of real-world LLIs and their corresponding daytime counterparts for the same scenes is difficult and challenging. To counter this problem, most techniques utilize synthetic LLIs produced from NLIs using light correction and noise induction techniques like gamma correction and Gaussian noise. However, synthetic LLIs do not always accurately represent real world low-light conditions, which usually encompass non-linear and spatially varying light conditions and noise levels, and are difficult to simulate mathematically. In an attempt to ease the restriction of paired or unpaired training labels and counter the synthetic LLIs performance problem, one recent approach by Guo et al. (2020) redefines the LLI enhancement task from an image-to-image learning task where the enhanced image is the final output of the network, to an image-to-curve estimation where light curves are learned and applied to enhance the image. The model achieves a good performance, thus opening new horizons for formulating the enhancement task. Another study by Jiang et al. (2021) describes a GAN-based unsupervised learning method (i.e., EnlightenGAN), performing enhancement without the need for training pairs or the LLIs' daytime counterparts. The latter achieves good performance levels and shows a lot of promise since unpaired image datasets are much easier to come by compared with paired ones. Another study by Khan et al. (2021c) also attempts to counter this challenge by proposing FSDG-Net: a Few Shot Divide and Glow network designed for end-to-end training from the correlation consistency of the input image decomposition itself using only few training samples. Experiments under very low exposure and ill-light conditions show good enhancement results, highlighting the potential of this solution.

Challenge 2: Most of the approaches tend to struggle whenever low-quality, noisy, or very dark images are considered during enhancement. This underlines the need for a proper understanding and modeling of the quality and noise elements in an image when conducting image enhancement or when designing a new LLI enhancement approach.

Challenge 3: Most existing techniques are developed as standalone solutions aiming to improve the illumination and the quality of LLIs. Yet, the latter's impact on high-level computer vision tasks like object detection and classification remains uncertain, where high-level image features might be distorted or lost during the enhancement task, thus leading to reduced or non-improving end-to-end performance.

Table 2. Characteristics of DL-based LLI enhancement models

(a) Encoder-decoder and CNN based models

Model	Description	Evaluation Metrics[2]	Loss functions[2]	Training Datasets
LLNet (Lore et al., 2017)	- First application of DL to enhance LLIs - Uses a sparse stacked denoising autoencoder for contrast enhancement and denoising	PSNR & SSIM (Wang et al., 2004)	L2	Synthetic based on CVG-UGR database[3]
LLCNN (Tao et al., 2017)	- Uses CNN modules based on inception modules and residual connections	PSNR, SSIM, LOE (Wang et al., 2013) & SNM (Hojatollah & Wang, 2013)	SSIM	Synthetic based on CVG-UGR database[3]
LL-RefineNet (Jiang et al., 2018)	- Uses two symmetrical paths: forward to extract high level features and backward to fuse and refine with low level features	PSNR, SSIM, & Root-MSE	L1 & L2	Synthetic based CVG-UGR database[3]
HDR-Net (Gharbi et al., 2017)	- Performs real time enhancement using a deep bilateral CNN which processes images in their low-resolution version	PSNR	L2	MIT-Adobe FiveK (Bychkovsky et al., 2011)
SID (Chen et al., 2018)	- Utilizes fully convolutional networks to enhance and denoise sensor raw data	PSNR & SSIM	L1	SID (Chen et al., 2018)
LRCNN (Xu et al., 2018)	- Uses a deep residual convolutional encoder-decoder along with fully connected layers for contrast enhancement and denoising	PSNR & SSIM	L2	Synthetic based on CVG-UGR database[3]
GladNet (Wang et al., 2018)	- Uses an encoder-decoder to estimate illumination and a CNN for content reconstruction	-----------	L1	Synthetic based on RAISE (Dang-Nguyen et al., 2015)

(b) Retinex theory based models

Model	Description	Evaluation Metrics	Loss functions	Training Datasets
MSR-Net (Shen et al., 2017)	- Uses Retinex theory to construct a CNN network that learns a mapping from dark to bright images	SSIM, NIQE (Mittal et al., 2013) & DE (Amigó et al., 2007)	L2	Synthetic based on UCID (Schaefer and Stich, 2003), BSD (Arbelaez et al., 2011), and Google images
RetinexNet (Wei et al., 2018)	- Decomposes the image into its reflectance and illumination components, then performs enhancement and denoising	------------	Reconstruction, IR, & IS	Real from LOL & synthetic based on RAISE (Dang-Nguyen et al., 2015)
LightenNet (Li et al., 2018)	- Learns an image to illumination map translation through a CNN	MSE, PSNR & SSIM	L2	Synthetic based on 600 pairs of normal & Retinex-darkened images
KinD (Zhang et al., 2019)	- Decomposes the image into reflectance and illumination components, clears reflectance degradations, and adjusts illumination	PSNR, SSIM, LOE, & NIQE	Based on: L1, L2, & SSIM	LOL (Wei et al., 2018)
DeepUPE (Wang et al., 2019)	- Learns an image-to-illumination mapping, then applies the Retinex model to enhance the image	PSNR & SSIM	Reconstruction, smoothness, & color	3000 pairs of underexposed & expert retouched images
FSDG-Net (Khan et al., 2021c)	- Comprises MID-Net and Glow-Net - Learns from the correlation consistency of image decomposition itself - Optimizes total variation operator	PSNR, SSIM, NIQE, and LOE	TV and L1 based losses	D3I (880 images), LOL (Wei et al., 2018), & EGAN dataset (Jiang et al., 2021)

(c) Fusion based models

Model	Description	Evaluation Metrics	Loss functions	Training Datasets
MBLLEN (Lv et al., 2018)	- Extracts features at every layer of the CNN, then enhances them via an encoder-decoder, and finally fuses them into an enhanced image	PSNR, SSIM, AB (Chen et al., 2006), VIF (Sheikh and Bovik, 2006), LOE, & TMQI (Hojatollah & Wang, 2013)	Structure, context & regional	Synthetic based on Pascal VOC (Everingham et al., 2012)
ACA-Net (Shin et al., 2018)	- Utilizes context aggregation networks to aggregate the global context of the image	PSNR & SSIM	Reconstruction & adversarial	Synthetic based on AVA (Murray et al., 2012)
DFN (Cheng et al., 2019)	- Combines features extracted using an encoder decoder from images derived by traditional methods	MSE, PSNR, SSIM, & NIQE	L1 & L2	Synthetic based on datasets from (Gu et al., 2016; Ma et al., 2017)
(Wang et al., 2019)	- Utilizes attention-based modules to enhance important features and suppress non-vital features	PSNR	PSNR	SID (Chen et al., 2018), S7ISP (Schwartz et al., 2019), & synthetic based on Pascal VOC (Everingham et al., 2012)
(Lv et al., 2020)	- Produces two attention maps to guide exposure enhancement and denoising, and performs enhancement using a multi-branch CNN	PSNR, SSIM, VIF, LOE, TMQI, AB, & LPIPS (Zhang et al., 2018)	L1, L2, bright, structure, perceptual, & regional	Synthetic based on publicly available datasets: Pascal VOC (Everingham et al., 2012), MS COCO (Lin et al., 2014), (Grubinger et al., 2006; Bileschi, 2006)
(Ren et al., 2019)	- Utilizes an RNN to learn salient edge contents and an encoder-decoder to learn the global contents	PSNR & SSIM	L2, perceptual, & adversarial	MIT-Adobe FiveK (Bychkovsky et al., 2011)
(Xiang et al., 2019)	- Uses a multi-branch encoder-decoder supplemented by ConvLSTM module	PSNR & SSIM	L1 & SSIM	LOL (Wei et al., 2018) & synthetic based on RAISE (Dang-Nguyen et al., 2015)

[2] PSNR: Peak Signal to Noise Ratio, SSIM: Structural Similarity Index, LOE: Lightness Order Error, SNM: Structure Natural Measure, NIQE: Natural Image Quality Evaluator, DE: Discrete Entropy, IR: Invariable Reflectance, IS: Illumination Smoothness, AB: Average Brightness, VIF: Visual Information Fidelity, TV: Total Variation, MSSIM: Multiscale SSIM, IQA: Image Quality Assessment, OAV: Open Available Vehicle, IWSSIM: Information-Weighted SSIM, PI: Perpetual Index, SC: Spatial Consistency, EC: Exposure Control, CC: Color Constancy, GE: Gray Entropy, CE: Color Entropy, GMI: Gray Mean Illumination, and GMG: Gray Mean Gradient.

[3] http://decsai.ugr.es/cvg/dbimagenes/

(d) GAN based models

Model	Description	Evaluation Metrics	Loss functions	Training Datasets
(Meng et al., 2019)	- Utilizes an encoder-decoder with a fusion network for the generator	Cosine similarity	Adversarial, perceptual, & TV	OAV dataset (Milford and Wyeth, 2012)
(Hua and Xia, 2018)	- Uses GANs joint with an image quality assessment network to improve visual quality	PSNR, SSIM, MSSSIM (Wang et al., 2003) & IWSSIM (Wang and Li, 2011)	Adversarial, content, TV & IQA	Synthetic based on General100 (Dong et al., 2016), Sun-Hayes80 (Sun and Hays, 2012) & Urbanal100 (Huang et al., 2015)
LowLightGAN (Kim et al., 2019)	- Uses spectral normalization to stabilize the training, and produces its dataset based on local illumination synthesis	NIQE	Adversarial, perceptual, color, & TV	Synthetic based on DIV2K (Agustsson and Timofte, 2017)
(Yangming et al., 2019)	- Combines Retinex Theory and GANs for image decomposition then enhancement	MSE, PSNR & SSIM	Regularization, adversarial, smooth L1, reconstruction, decomposition, enhancement, & MSSSIM	CSID (Chen et al., 2018)
RDGAN (Wang et al., 2019)	- Learns to decompose the image into reflectance and illumination, then fuses them into a final enhanced image	PSNR & FSIMc (Zhang et al., 2011)	Multi term decomposition, content & adversarial	SICE (Cai et al., 2018)
DPE (Chen et al., 2018)	- Utilizes paired and unpaired training settings based on GANs for enhancement	PSNR & SSIM	L2 & adversarial	MIT-Adobe 5K (Bychkovsky et al., 2011), & HDR images selected from Flickr images
EnlightenGAN (Jiang et al., 2021)	- Performs unsupervised learning, using a U-Net (Ronneberger et al. 2015) as the backbone for the generator, and includes global and local relativistic discriminators to handle spatially varying light conditions	NIQE	Non-reference: self-feature preserving, & adversarial	Collected from LOL (Wei et al., 2018), RAISE (Dang-Nguyen et al., 2015), & HDR sources (Gharbi et al., 2017; Kalantari and Ramamoorth, 2017)

(e) Zero reference models

ZeroDCE (Guo et al., 2020)	- Learns image-to-light enhancement curve mappings through a lightweight CNN model - Does not require paired or unpaired training data	PSNR, SSIM, MAE & PI (Blau and Michaeli, 2018)	Non-reference: SC, EC, CC & IS	360 multi-exposure sequences from part1 of SICE (Cai et al., 2018)
(Zhang et al., 2020)	- Introduces a reference-less loss function designed based on the Retinex model and entropy theory to train a self-supervised CNN model - Does not require paired or unpaired training data	GE, CE, GMI, GMG, PSNR, SSIM, LOE, & NIQE	Non-reference: reconstruction, reflectance, & IS	LOL (Wei et al., 2018)
SCI (Ma et al., 2022)	- Learns illumination map using cascaded stages of intermediate representations - Relies on self-calibrated module to ensure convergence between representations	PSNR, SSIM, DE, EME (Agaian et al., 2007), LOE & NIQE	Non-reference: fidelity and smoothing	-----------

Challenge 4: It is difficult to fairly compare most existing models for two main reasons: i) lack of a large standard dataset of paired LLIs/NLIs that are taken from real-world scenes and represent various low-light conditions, and ii) lack of a (set of) common and standard metric(s) that can accurately evaluate the visual perception of enhanced image quality. As can be seen in Table 2, different datasets and evaluation metrics are used to train and evaluate the visual performance of different enhancement models.

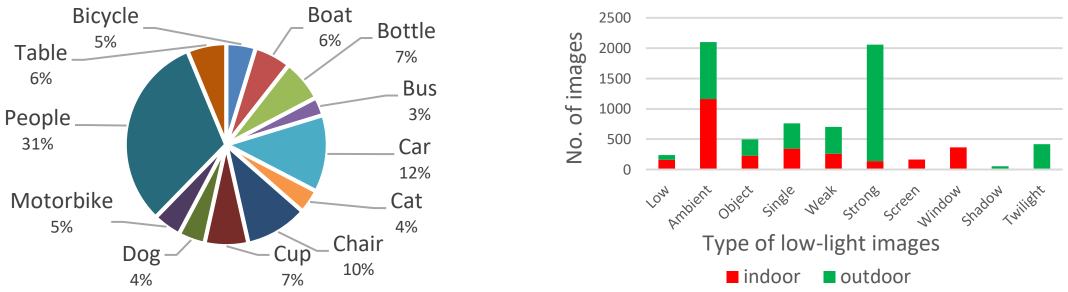
In the following empirical study, we further discuss the above challenges aiming to acquire a better understanding of the issues at stake and shed light on possible future directions.

4. Overview of Experimental Study

One of the main challenges facing LLI enhancement is the lack of common benchmarks and metrics for empirical evaluation. In this section, we describe the test data, experiments, and evaluation metrics that we adopt in our study.

4.1. Test Data

We use two well-known datasets to conduct our empirical evaluation: ExDark (Loh and Chan, 2019) and LOL (Wei et al., 2018). ExDark consists of 7,363 LLIs captured in real-world natural scene low-light environments, and contains 12 different object classes like people, cats, dogs, bicycles, etc. (Fig. 2a). Every instance of the 12 classes is associated with a bounding box annotation making the dataset applicable for training and evaluation on object detection and classification models. In addition, the dataset is split among 10 types of low-light conditions found in indoor and outdoor environments, varying from extremely dark images to images with spatially varying illumination depending on the location and the presence of light sources (Fig. 2b).



(a) Object occurrences in dataset

(b) Distribution of images on illumination types

Fig. 2. ExDark statistics reported from (Loh and Chan, 2019)

LOL (Low-light) (Wei et al., 2018) is made of 500 LLI/NLI pairs. The NLIs refer to a variety of real scenes taken in houses, campuses, clubs, etc. Yet, most of their LLI counterparts are created by changing the camera exposure and ISO sensitivity of the image sensor in order to simulate low-light conditions and thus they do not represent real low-light environments (Loh and Chan, 2019). Hence, we refer to LOL as a quasi-synthetic dataset. Note that LOL images do not contain moving objects (such as people, animals, and vehicles) as the image pairs require exact position matching between LLI/NLI pairs (sample LOL image pairs are shown in Fig. 3).

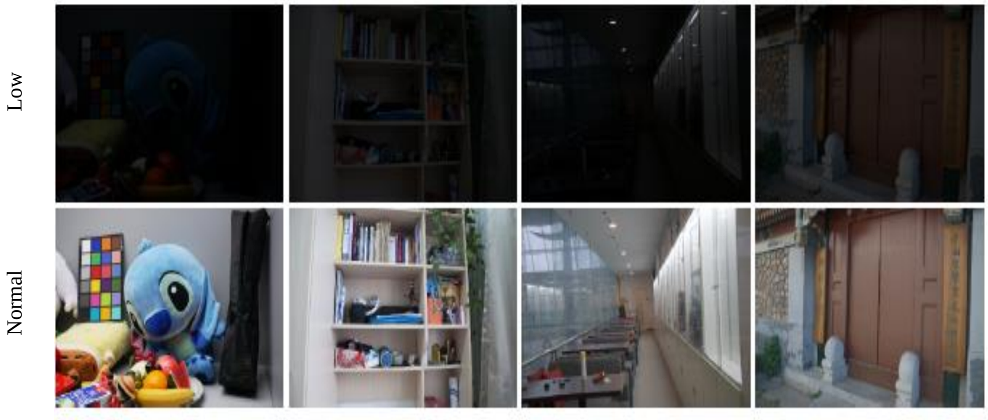


Fig. 3. Sample pairs of LLIs/NLIs from the LOL dataset (Wei et al., 2018)

4.2. Experiments and Metrics

Our empirical evaluation consists of three main experiments: i) visual and perceptual quality evaluation, ii) detection and classification quality evaluation, and iii) feature analysis.

4.2.1. Experiment 1 – Visual and Perceptual Quality

In this experiment, we perform an image quality assessment (IQA) that aims at evaluating whether an image is visually pleasing and how it is visually perceived. Image quality refers to the different visual attributes of the image and focuses on the perceptual assessment of viewers. IQA methods are generally either i) quantitative: based on objective evaluation metrics, or ii) qualitative: based on the human perception of visual quality. In this study, we conduct both quantitative and qualitative evaluations, by comparing the visual quality achieved by 10 of the recent DL-based LLI enhancement models.

Quantitative comparison: We evaluate the enhancement models against four objective evaluation metrics commonly used in the literature: i) Natural Image Quality Evaluator (*NIQE*) (Mittal et al., 2013), ii) Blind/Referenceless Image Spatial Quality Evaluator (*BRISQUE*) (Mittal et al., 2012), iii) Structural Similarity Index (*SSIM*) (Wang et al., 2004) and iv) Peak Signal to Noise Ratio (*PSNR*).

NIQE (Mittal et al., 2013) is a *non-reference* metric or "blind" evaluation metric in which only the LLIs are available for assessment. It measures the deviations from statistical regularities seen in natural images without training on human rated distorted images or even exposure to distorted images. The quality of the test image represents the distance between a multivariate Gaussian (MVG) fit of the natural scene statistic (NSS) features derived from the test image, and a MVG model of the quality aware features extracted from a corpus of natural images.

BRISQUE (Mittal et al., 2012) is also a *non-reference* evaluation metric. It belongs to a class of opinion-aware metrics which evaluate the image based on models trained on databases of human rated distorted images and associated subjective opinion scores. In *BRISQUE*, the extracted features are derived based on a spatial natural scene statistical model. Then, a mapping is learned between the feature space and human based quality scores using a regression module, namely a support vector machine regressor (SVM-R) (Schölkopf et al., 2000).

SSIM (Wang et al., 2004) is a *full reference* metric in which a known reference image is needed for assessment. It measures the structural similarity between images based on independent comparisons of their luminance, contrast, and structure features. Given a ground truth image *x* with *N* pixels and maximum pixel value *L*, and given the corresponding enhanced image *y*, a simplified version of *SSIM* is defined as follows (Wang et al., 2004):

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C1)(2\sigma_{xy} + C2)}{(\mu_x{}^2 + \mu_y{}^2 + C1)(\sigma_x{}^2 + \sigma_y{}^2 + C2)} \tag{3}$$

where $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\sigma_x = (\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2)^{1/2}$, $\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$, $C1 = (k_1 L)^2$, and $C2 = (k_2 L)^2$ are constants for avoiding instability, with k1 << 1 and k2 << 1.

PSNR is another commonly used *full reference* evaluation metric. It is defined using the maximum pixel value (denoted as *L*) and the mean squared error (MSE or L2 loss) between images. Given a ground truth image x with *N* pixels and the corresponding enhanced image *y*, the *PSNR* between *x* and *y* is defined as follows:

$$PSNR(x, y) = 10 \times \log_{10}\left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}(x(i) - y(i))^2}\right) \tag{4}$$

In addition to the above objective metrics, we also evaluate the *noise level* in the enhanced images to acquire a complete viewpoint of the enhancement quality achieved by the models. We follow the approach proposed in (Liu et al., 2012) which relies on a patch-based noise level estimation algorithm. The algorithm selects weak textured patches from a single noisy image based on the gradients of the patches and their statistics. Then it estimates the noise level from the selected patches using principal component analysis.

Finally, we note that our methodology to conduct the experiments relies on commonly used IQA metrics and LLI datasets available in the literature. Yet, other IQA metrics and LLI datasets can be used, as mentioned in Table 2. Here, we emphasize and re-iterate the difficulty in conducting comparative evaluations due to the lack of golden standard metrics and datasets for LLI enhancement evaluation, and highlight the need to create a uniform reference benchmark for comparing LLI solutions (cf. challenge #4 in Section 3.6).

Qualitative comparison: In addition to the quantitative study, we also perform a qualitative evaluation to assess the human visual perception of images enhanced by the 10 models used in this experiment. To do so, we randomly select 20 LLIs from our test data, i.e., 10 from each dataset (ExDark and LOL), and display them along with their enhanced counterparts in two dedicated surveys (for the LOL survey, we also display the corresponding NLIs)[4]. Responders are asked to rate each image considering six visual IQA criteria including: i) level of illumination, ii) level of exposure (over/under-exposed regions), iii) level of noise, iv) color deviations, v) clearness of contents and details, and vi) overall beauty. A total of 32 testers (senior computer engineering and master's students) were invited to contribute in the experiment, where 16 testers participated in each survey and independently rated every enhancement model on an integer scale from 1 to 10 (i.e., worst to best). A total of 1,600 responses were collected for each dataset, with every model receiving 160 rating scores. The ratings are aggregated for every enhancement model to evaluate its overall visual perceptual quality.

4.2.2. Experiment 2 – Detection and Classification Quality

In this experiment, we compare the performance achieved by 4 different object detection and classification models applied on the enhanced images from ExDark dataset using the 10 LLI enhancement methods considered in the previous experiment. We utilize mean average precision (*mAP*) as a commonly used metric to assess object detection and classification quality. For each object class, we generate the corresponding precision-recall (*P-R*) curve and compute the average precision (*AP*) per class from the area covered under the *P-R* curve. We then compute *mAP* for the object detection model as the average of the *AP* scores calculated for all the classes.

4.2.3. Experiment 3 – Feature Analysis

In this experiment, we compare the feature maps extracted from the LLIs, NLIs, and enhanced images from the LOL dataset using one of the object detection models from Experiment 2. A feature map is an $m\times n$ matrix which represents the output of a filter applied to a layer of the object detection model. A layer in a DL-based model usually consists of a sequence of feature maps. In this experiment, we consider the feature maps from three sample layers of the detection model: i) a sequence of large (e.g., 64×64 cell) maps from one of the layers belonging to the model's backbone, ii) a sequence of smaller (e.g., 16×16 cell) maps from an intermediary layer, and iii) a sequence of minimal size (e.g, 1×1 cell) maps from the model's last layer. To our knowledge, this is the first quantitative feature map evaluation study of its kind in the literature. We introduce two new metrics to compare feature maps: i) Feature Map Matrix Similarity (*FMMS*), and ii) top-*N* Active Feature Map Similarity (*topN-AFMS*). *FMMS* computes the cosine similarity measure between the feature maps of two (sets of) images at a given layer of the DL model, highlighting overall image feature similarity. More formally, given two images *x* and *y* whose feature maps are extracted at layer *n* of the DL model:

$$FFMS(x,y) = \frac{\sum_{i=1\ldots|n|} Sim_{Cosine}(F_i^x, F_i^y)}{|n|} \in [0,1] \tag{5}$$

where F_i^x and F_i^y are the i^{th} feature maps of images *x* and *y*, $|n|$ the number of feature maps at layer *n*, and Sim_{cosine} the legacy cosine matrix similarity measure[5]:

$$Sim_{Cosine}(F_i, F_j) = \frac{\sum_q \sum_r w_i(q,r) \times w_j(q,r)}{\sqrt{\sum_q \sum_r w_i(q,r)^2 \times \sum_q \sum_r w_j(q,r)^2}} \in [0,1] \tag{6}$$

[4] ExDark: https://cutt.ly/0fN4evQ, and LOL: https://cutt.ly/TfN4r6G

[5] We adopt the cosine measure due to its common usage in the literature (McGill, 1983), yet other vector or matrix similarity measures could have been used such as Pearson Correlation Coefficient or Dice.

where $w_i(q, r)$ is the feature map (matrix) F_i position at coordinates q and r. Note that *FMMS* can be extended to compare two sets of pair-wise matching images (e.g., comparing sets of LLI/NLI, LLI/enhanced, or NLI/enhanced image pairs) by computing the similarity between every matching pair and then averaging over the total number of image pairs.

As for *topN-AFMS*, it compares the most active feature maps between two sets of pair-wise matching images, in order to help describe the behavior of a detection model and its response activity against the fed images. Identifying the most active feature maps gives insight into the features that might be most impactful on object detection and classification quality. Given two sets of pair-wise matching images $X=\{x_1, \dots, x_t\}$ and $Y=\{y_1, \dots, y_t\}$ where doublet (x_i, y_i) designates a matching pair (e.g., LLI/NLI, LLI/enhanced, or NLI/enhanced), and given the images' feature maps extracted at layer n of the DL model, we produce two vectors $V_X = < w_X(1), \dots, w_X(|n|) >$ and $V_Y = < w_Y(1), \dots, w_Y(|n|) >$ of size $|n|$ each, where weights $w_X(i)$ and $w_Y(i)$ designate the number of times feature map i at layer n occurs among the top-active feature maps (based on their average) in image set X and Y respectively. For instance, $w_X(i)=10$ means that feature map i has been identified 10 times (i.e., in 10 different images of set X) as one of the top active feature maps at layer n. Consequently, computing *topN-AFMS* between image sets X and Y comes down to computing the similarity between their vectors:

$$\text{top}N\text{-AFMS}(\mathrm{X},\mathrm{Y}) = \mathrm{Sim}_{\mathrm{Cosine}}(\mathrm{V_X},\mathrm{V_Y}) = \frac{\sum_{\mathrm{i}=1\dots|n|} w_X(\mathrm{i}) \times w_Y(\mathrm{i})}{\sqrt{\sum_{r=1\dots|n|} w_X(\mathrm{i})^2 \times \sum_{r=1\dots|n|} w_Y(\mathrm{i})^2}} \in [0,1] \tag{7}$$

where N is the number of most active feature maps at a certain layer of the DL model (e.g., we consider N=16 and compute the *top16-AFMS* in our empirical study, cf. Section 7.1).

In addition to computing *FFMS* and *topN-AFMS*, we utilize the occlusion experiment proposed in (Zeiler and Fergus, 2014), where a black square is used to mask particular regions of an image while monitoring the output of the object detection model. The black square is slided over all the regions of the image allowing to produce a heatmap describing object detection confidence scores (in case of a detection – zero scores are produced otherwise). The significance of the experiment lays in the fact that the output of the object detection model should not change when the regions that are not so important for detection are occluded, and should vanish when the regions responsible of the detection are occluded. The occlusion experiment is applied on images containing a single object. If an image has a lot of regions that result in a misdetection if occluded, then we say the image holds weak features allowing to easily misdetect its object. Contrarily, if an image has no specific region that causes misdetection when occluded, then the image maintains strong features allowing to detect its object despite occlusion. While the authors in (Zeiler and Fergus, 2014) describe the occlusion experiment, yet they do not define a quantitative approach to evaluate its results. Here, we introduce an objective metric: Occlusion based Average Misdetection Regions (*OAMR*) that quantifies the average number of regions contributing to misdetecting objects in a set of images. More formally, given a set of images $X=\{x_1, \dots, x_n\}$ with n images of same size and a fixed size black box sliding over all the regions of the image then:

$$OAMR(X) = \frac{\sum_{i=1\dots n} C_{x_i}}{n} \tag{8}$$

where $C_{x_i}$is the count of regions contributing to a misdetection in image x_i. A low *OAMR* indicates that a small number of regions causes misdetections, meaning that the images mostly contain strong features contributing to high object detection quality. A high *OAMR* indicates that many regions cause misdetections, and thus the images hold weak features leading to low object detection quality. In short, high quality LLI enhancement models would minimize *OAMR*.

5. Experiment 1: Perceptual and Visual Quality

In this section, we present quantitative and qualitative evaluations of the performance achieved by 10 recent DL-based LLI enhancement models, namely: RetinexNet[6] (Wei et al., 2018), GladNet[7] (Wang et al., 2018), LLNet[8] (Lore et al., 2017), LightenNet[9] (Li et al., 2018), DPE[10] (Chen et al., 2018), EnlightenGAN[11] (Jiang et al., 2021), MBLLEN[12] (Lv et al., 2018), DeepUPE[13] (Wang et al., 2019), RDGAN[14] (Wang et al., 2019), and ZeroDCE[15] (Guo et al., 2020). We run the latter on both ExDark and LOL datasets using the models' pre-trained weights and author-recommended configurations which are publicly available online.

We first perform a runtime analysis for the 10 enhancement solutions considered in our experiment. Table 3 shows the runtime for each solution averaged over 100 images of size 600×700×3 from the ExDark dataset. Measurements were conducted using a GPU (Nvidia Tesla K80), except for LightenNet (Li et al., 2018) which was evaluated on a CPU (Intel I7 6700) since its original code is written in MATLAB. Results in Table 3 show that ZeroDCE (Guo et al., 2020) has superior runtime compared with its counterparts, reflecting its real time enhancement capability. Most remaining models produce comparable runtime results except for MBLLEN, LLNET and LightenNet which show inferior runtime and rank at the bottom of the list. LightenNet produces the worst runtime results, which is probably due to its MATLAB implementation and CPU execution.

Table 3. Runtime comparison for the 10 DL LLI enhancement models evaluated in our study, ranked from best to worst

Approach	**Runtime (sec) ↓**	**Platform**
ZeroDCE	0.0024	PyTorch (GPU)
EnlightenGAN	0.0116	PyTorch (GPU)
DeepUPE	0.0190	TensorFlow (GPU)
RetinexNet	0.1960	TensorFlow (GPU)
DPE	0.2337	TensorFlow (GPU)
GladNet	0.2363	TensorFlow (GPU)
RDGAN	0.2841	TensorFlow (GPU)
MBLLEN	0.6395	TensorFlow (GPU)
LLNet	1.0521	Theano (GPU)
LightenNet	5.2607	MATLAB (CPU)

5.1. Quantitative Comparison

To perform a quantitative evaluation, we process the results produced by each of the mentioned DL models through four commonly used metrics in the literature: i) Natural Image Quality Evaluator (*NIQE*) (Mittal et al., 2013), ii) Blind/Reference-less Image Spatial Quality Evaluator (*BRISQUE*) (Mittal et al., 2012), iii) Structural Similarity Index (*SSIM*) (Wang et al., 2004) and iv) Peak Signal to Noise Ratio (*PSNR*) (cf. Section 4.2.1). We use the first two metrics, i.e., *NIQE* and *BRISQUE*, to evaluate the enhanced images from the ExDark and LOL datasets as stand-alone images

[6] https://github.com/weichen582/RetinexNet
[7] https://github.com/weichen582/GLADNet
[8] https://github.com/kglore/llnet_color
[9] https://li-chongyi.github.io/sub_projects.html
[10] https://github.com/UtopiaHu/Deep-Photo-Enhancer
[11] https://github.com/TAMU-VITA/EnlightenGAN
[12] https://github.com/Lvfeifan/MBLLEN
[13] https://github.com/wangruixing/DeepUPE
[14] https://github.com/WangJY06/RDGAN
[15] https://github.com/Li-Chongyi/Zero-DCE

without referring to their NLI counterparts. We use the third and fourth metrics to evaluate the enhanced images from the LOL dataset against their reference NLI counterparts. We also evaluate noise levels in the enhanced images[16] as an added indicator of the enhancement quality achieved by the models. In addition to the 10 models being evaluated, we also provide the scores obtained for the original LLIs, which we use as a reference to compare with the latter. Models producing *NIQE* and *BRISQUE* scores that are lower/higher than the original LLI scores are considered to be better/worse in improving the visual quality of the LLIs, and models producing *PSNR* and *SSIM* scores that are higher/lower than the original LLI scores are considered to be better/worse in improving the visual quality of the LLIs. Results for the ExDark and LOL datasets are provided in Table 4 and Table 5 and samples are visualized in Fig. 5 and Fig. 6. We further report in appendices A and B the boxplots for the distribution of the metric measures obtained for the original and enhanced image datasets, along with standard deviations and number of outliers. The boxplots are relatively squeezed and short indicating high alignment among the measures. This reflects the consistent enhancement and noise effect applied by each of the enhancement solutions on the original images. Moreover, we note that the standard deviations are significantly small compared with the corresponding mean values, indicating that the mean measure is a good representative of the metrics' distribution. In addition, the number of outliers reported by the boxplots is small compared with the total number of measures (i.e., 7,363 for ExDark and 500 for LOL) emphasizing the highly aligned measures around the median of the distributions.

Table 4. Results for the ExDark dataset, ranked from best (#1) to worst (#10) following each of the metrics (red color refers to the best score and green to the second best for every metric)

(a) Results ranked following *NIQE* ↓

Approach	**Rank**	***NIQE***	***BRISQUE***	**Noise**
MBLLEN	1	**3.26**	**25.75**	**19.95**
EnlightenGAN	2	**3.52**	26.33	20.07
DPE	3	3.64	29.04	20.108
LightenNet	4	3.6573	27.52	20.16
DeepUPE	5	3.6575	28.09	20.106
Original LLIs	--	3.68	30.51	20.00
GladNet	6	3.76	28.21	20.14
RDGAN	7	3.87	**26.24**	20.15
LLNet	8	3.99	33.42	**19.77**
ZeroDCE	9	4.05	30.49	20.28
RetinexNet	10	4.26	31.71	20.101

(b) Results ranked following *BRISQUE* ↓

Approach	**Rank**	***NIQE***	***BRISQUE***	**Noise**
MBLLEN	1	**3.26**	**25.75**	**19.95**
RDGAN	2	3.87	**26.24**	20.15
EnlightenGAN	3	**3.52**	26.33	20.07
LightenNet	4	3.6573	27.52	20.16
DeepUPE	5	3.6575	28.09	20.106
GladNet	6	3.76	28.21	20.14
DPE	7	3.64	29.04	20.108
ZeroDCE	8	4.05	30.49	20.28
Original LLIs	---	3.68	30.51	20.00
RetinexNet	9	4.26	31.71	20.101
LLNet	10	3.99	33.42	**19.77**

(c) Results ranked following *Noise Level* ↓

Approach	**Rank**	***NIQE***	***BRISQUE***	***Noise***
LLNet	1	3.9	33.42	**19.77**
MBLLEN	2	**3.26**	**25.75**	**19.95**
Original LLIs	---	3.68	30.51	20.00
EnlightenGAN	3	**3.52**	26.33	20.07
RetinexNet	4	4.26	31.71	20.101
DeepUPE	5	3.65	28.09	20.106
DPE	6	3.64	29.04	20.108
GladNet	7	3.76	28.21	20.14
RDGAN	8	3.87	**26.24**	20.15
LightenNet	9	3.65	27.52	20.16
ZeroDCE	10	4.05	30.49	20.28

Based on the results in Table 4 and Table 5, we highlight the following observations:

- **Results of the *NIQE* and *BRISQUE* metrics do not seem correlated**: Models that perform well following the first metric might perform poorly with the second. Considering the ExDark dataset for instance, DPE has the third best *NIQE* score while showing the fourth worst *BRISQUE* score. Also, RDGAN achieves the second best *BRISQUE* score while showing the fourth worst *NIQE* score. Similarly, for the LOL dataset, LLNet which is the best following the *NIQE* metric produces the worst *BRISQUE* score.
- **LOL dataset results show discrepancies among *NIQE*, *BRISQUE*, *SSIM*, and *PSNR* metrics**: For instance, GladNet is ranked as the fourth worst model following *NIQE* and *BRISQUE*, yet it shows the best *SSIM* and *PSNR* scores. Also, MBLLEN which achieves the best score for *BRISQUE* and the second best for *NIQE* comes only at the fifth place following *SSIM*.

[16] For noise level evaluation, we use: i) a random subset of 500 images from ExDark including 50 images from each of the 10 different lighting conditions, as well as ii) the whole LOL dataset.

- **Most of the models produce noise levels higher than the original LLIs, with a few exceptions:** Most enhancement models tend to amplify or integrate noise into the enhanced images, except for LLNet and MBLLEN with both ExDark and LOL. LLNet achieves the minimum noise levels as it tends to over-smooth image details. MBLLEN produces the second lowest noise levels and is consistent with the good scores achieved by both *NIQE* and *BRISQUE* metrics with both ExDark and LOL datasets thus highlighting its good enhancement performance. EnlightenGAN produces some of the best scores on the ExDark dataset and adds a minimal amount of noise. In contrast, ZeroDCE consistently shows the highest noise levels and produces some of the worst *NIQE* and *BRISQUE* scores, indicating that a high noise level tends to distort enhancement quality.

Table 5. Results for the LOL dataset, ordered from best (#1) to worst (#10) following each of the metrics (red color refers to the best score and green to the second best for every metric)

(a) Results ranked following *NIQE*↓

Approach	Rank	*NIQE*	*BRISQUE*	*SSIM*	*PSNR*	Noise
LLNet	1	**4.17**	33.03	**0.66**	**17.50**	**19.75**
MBLLEN	2	**4.22**	**20.38**	0.59	17.30	**19.99**
EnlightenGAN	3	4.97	24.41	0.60	16.25	20.15
RetinexNet	4	5.30	**24.19**	0.57	16.23	20.55
Original LLIs	---	6.03	24.87	0.16	7.74	19.99
DPE	5	6.64	24.51	0.371	9.41	20.09
RDGAN	6	7.04	27.85	0.62	14.97	20.49
GladNet	7	7.23	28.67	**0.67**	**19.26**	20.86
LightenNet	8	7.68	28.81	0.370	10.13	20.25
DeepUPE	9	7.81	27.91	0.39	10.57	20.11
ZeroDCE	10	8.54	31.80	0.54	14.16	21.35

(b) Results ranked following *BRISQUE*↓

Approach	Rank	*NIQE*	*BRISQUE*	*SSIM*	*PSNR*	Noise
MBLLEN	1	**4.22**	**20.38**	0.59	17.30	**19.99**
RetinexNet	2	5.30	**24.19**	0.57	16.23	20.55
EnlightenGAN	3	4.97	24.41	0.60	16.25	20.15
DPE	4	6.64	24.51	0.371	9.41	20.09
Original LLIs	---	6.03	24.87	0.16	7.74	19.99
RDGAN	5	7.04	27.85	0.62	14.97	20.49
DeepUPE	6	7.81	27.91	0.39	10.57	20.11
GladNet	7	7.23	28.67	**0.67**	**19.26**	20.86
LightenNet	8	7.68	28.81	0.370	10.13	20.25
ZeroDCE	9	8.54	31.80	0.54	14.16	21.35
LLNet	10	**4.17**	33.03	**0.66**	**17.50**	**19.75**

(c) Results ranked following *SSIM* ↑

Approach	Rank	*NIQE*	*BRISQUE*	*SSIM*	*PSNR*	Noise
GladNet	1	7.23	28.67	**0.67**	**19.26**	20.86
LLNet	2	**4.17**	33.03	**0.66**	**17.50**	**19.75**
RDGAN	3	7.04	27.85	0.62	14.97	20.49
EnlightenGAN	4	4.97	24.41	0.60	16.25	20.15
MBLLEN	5	**4.22**	**20.38**	0.59	17.30	**19.99**
RetinexNet	6	5.30	**24.19**	0.57	16.23	20.55
ZeroDCE	7	8.54	31.80	0.54	14.16	21.35
DeepUPE	8	7.81	27.91	0.39	10.57	20.11
DPE	9	6.64	24.51	0.371	9.41	20.09
LightenNet	10	7.68	28.81	0.370	10.13	20.25
Original LLIs	---	6.03	24.87	0.16	7.74	19.99

(d) Results ranked following *PSNR*↑

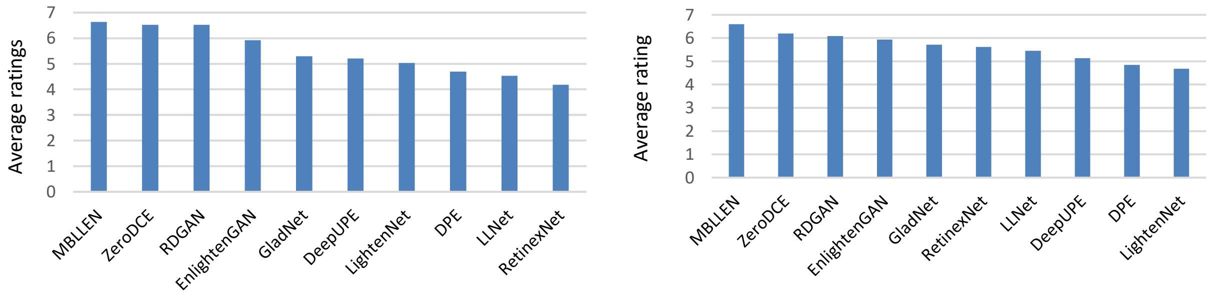
Approach	Rank	*NIQE*	*BRISQUE*	*SSIM*	*PSNR*	Noise
GladNet	1	7.23	28.67	**0.67**	**19.26**	20.86
LLNet	2	**4.17**	33.03	**0.66**	**17.50**	**19.75**
MBLLEN	3	**4.22**	**20.38**	0.59	17.30	**19.99**
EnlightenGAN	4	4.97	24.41	0.60	16.25	20.15
RetinexNet	5	5.30	**24.19**	0.57	16.23	20.55
RDGAN	6	7.04	27.85	0.62	14.97	20.49
ZeroDCE	7	8.54	31.80	0.54	14.16	21.35
DeepUPE	8	7.81	27.91	0.39	10.57	20.11
LightenNet	9	7.68	28.81	0.371	10.13	20.25
DPE	10	6.64	24.51	0.370	9.41	20.09
Original LLIs	---	6.03	24.87	0.16	7.74	19.99

(e) Results ranked following *Noise Level* ↓

Approach	Rank	*NIQE*	*BRISQUE*	*SSIM*	*PSNR*	Noise
LLNet	1	**4.17**	33.03	**0.66**	**17.50**	**19.75**
Original LLIs	---	6.03	24.87	0.16	7.74	19.99
MBLLEN	2	4.22	**20.38**	0.59	17.30	**19.99**
DPE	3	6.64	24.51	0.37	9.41	20.09
DeepUPE	4	7.81	27.91	0.39	10.57	20.11
EnlightenGAN	5	4.97	24.41	0.60	16.25	20.15
LightenNet	6	7.68	28.81	0.37	10.13	20.25
RDGAN	7	7.04	27.85	0.62	14.97	20.49
RetinexNet	8	5.30	24.19	0.57	16.23	20.55
GladNet	9	7.23	28.67	**0.67**	**19.26**	20.86
ZeroDCE	10	8.54	31.80	0.54	14.16	21.35

5.2. Qualitative Comparison

In addition to the quantitative study, we also perform a qualitative evaluation, where 32 participants were asked to rate samples of enhanced images from each of the two datasets used in our experiments (Fig. 5 and Fig. 6), providing their perception of the images' visual quality. Every sample image was independently rated on an integer scale from 1 to 10 (i.e., worst to best). Results for each enhancement model are compiled in Fig. 4.



(a) ExDark dataset scores

(b) LOL dataset scores

Fig. 4. Average tester rating scores compiled for every enhancement model, and ranked from best to worst

Based on the results in Fig. 4, we highlight the following observations:

- **Regarding the ExDark dataset**: MBLLEN tends to entirely illuminate the images to look visually pleasing and beautiful (e.g. bicycle and cat in Fig. 5a, c) and is ranked as the best enhancement model. Images enhanced by ZeroDCE and RDGAN show good illumination levels and well-preserved contents, and are ranked as second and third best models. EnlightenGAN tends to produce visually pleasing images with no over or under exposed regions, and is ranked as the fourth best model. GladNet tends to increase image illumination but shows some color deviation and noise, and is ranked as the fifth best model. DeepUPE, LightenNet and DPE add minimal touches on the images and tend to show low illumination levels. They are ranked at the sixth, seventh, and eighth positions, respectively. LLNet increases image illumination, yet it also tends to over-smooth certain image details (e.g., pedestrian street in Fig. 5a). It is ranked as the ninth and second last model. Finally, RetinexNet is ranked as the tenth and last model as it produces significant noise and tends to over-expose certain artifacts in the enhanced images (e.g., bicycle in Fig. 5a).
- **Regarding the LOL dataset**: MBLLEN produces visually pleasing images with vivid and natural colors and is ranked as the best model thus demonstrating its good enhancement quality. RDGAN, ZeroDCE, and EnlightenGAN show naturalistic colors with preserved contents and texture. They are ranked at the second, third, and fourth positions, respectively. GladNet sufficiently boosts image illumination but usually shows pale colors and tends to add noise. It is ranked at the fifth position. RetinexNet boosts image illumination while showing exposed artifacts. It is ranked at the sixth position. LLNet tends to highly smoothen image details while showing pale lighting, and is ranked at the seventh position. DeepUPE, DPE and LightenNet minimally enhance image illumination and tend to incorporate noise into the images. Together as a group, they are ranked as the three worst models.

5.3. Discussion

To sum up, we review and summarize the results of both quantitative and qualitative tests.

First, concerning the *quantitative evaluation metrics*: most metrics used in this study fail to produce model rankings which closely match the qualitative (human) evaluation rankings. For instance, *NIQE* results are compatible with the qualitative scores in certain aspects, by i) showing that MBLLEN achieves the best/second best enhancement quality for ExDark/LOL datasets, ii) showing that RetinexNet and LLNet are amongst the worst performing models when applied on the ExDark dataset, and iii) producing very close scores for DPE, LightenNet, and DeepUPE which only perform slight enhancement to the images of ExDark, in accordance with their sequential human rankings. However, *NIQE* does not show consistent results when it comes to capturing the illumination and noise components in the enhanced images. For instance, in the case of the ExDark dataset, ZeroDCE and RDGAN are ranked among the worst models following *NIQE* as they produce high noise levels (Table 4c). Yet, they are relatively better ranked by human testers, producing higher scores than DPE, LightenNet, and DeepUPE which have better *NIQE* scores and lower illumination levels. This might be due to the fact that the noise produced by ZeroDCE and RDGAN is not clearly apparent in the images and maybe visually overlooked by the users in favor of good illumination. Moreover, LLNet shows the best *NIQE* score while maintaining the lowest noise level for the LOL dataset, yet it exhibits the fourth worst human scores due to the over-smoothing and the exposed artifacts it produces. *BRISQUE* shows similar inconsistencies while quantifying image illumination. For example, ZeroDCE shows higher *BRISQUE* scores compared with DeepUPE, DPE, and LightenNet, and yet surpasses the latter models in terms of human tester ratings. This is probably due to the seemingly better illumination as perceived by most testers. In addition, all considered metrics fail to produce consistent rankings among themselves, suggesting the need to design more accurate objective metrics that behave in accordance with human visual perception.

Second, concerning the *best performing models*: EnlightenGAN is consistently ranked among the best enhancement models on both ExDark and LOL datasets. Although its good performance on LOL can be due to using it as part of the model's training dataset, yet its performance on ExDark proves its capability of generalizing to real world scenes. The results of this model highlight the potential of unsupervised GAN-based solutions in performing LLI enhancement. ZeroDCE is ranked as one of the best models following human ratings. It shows good illumination levels and preserves image contents, but tends to incorporate noise into the enhanced images (producing the highest noise levels for both datasets). The latter highlights the potential of ZeroDCE which reformulates the enhancement task using image-to-light curve estimation mapping, while eliminating the need for paired and unpaired training data. Also, the supervised MBLLEN model achieves some of the best quantitatively and qualitatively enhancement results on both ExDark and LOL datasets. This may be due to its large training dataset of synthetic LLIs (16,925 images) generated based on the Pascal VOC (Everingham et al., 2012) object detection and classification benchmark, allowing it to better generalize and handle real-world LLIs (namely those in ExDark and LOL). In addition, MBLLEN extracts and enhances the features at every layer of the used CNN model thus allowing global and local level feature enhancement.

Third, concerning the *noise element*: most enhancement models tend to incorporate significant noise into the enhanced images, thus distorting their quality. Notably, LLNet achieves minimal noise levels on both datasets, while sufficiently boosting image illumination. Its underlying Stacked Sparse Denoise Autoencoder (SSDA) (Lore et al. 2017) seems promising and could be effective if properly tuned and designed to maintain a good balance between noise suppression and over-smoothing. Nonetheless, we note that the noise factor and de-noising techniques need to be given special attention, especially that the present evaluation metrics do not simultaneously quantify illumination and noise levels. This might suggest the need for new and more robust metrics that are consistent with the humans' visual perception of enhanced image quality.

6. Experiment 2: Detection & Classification Quality

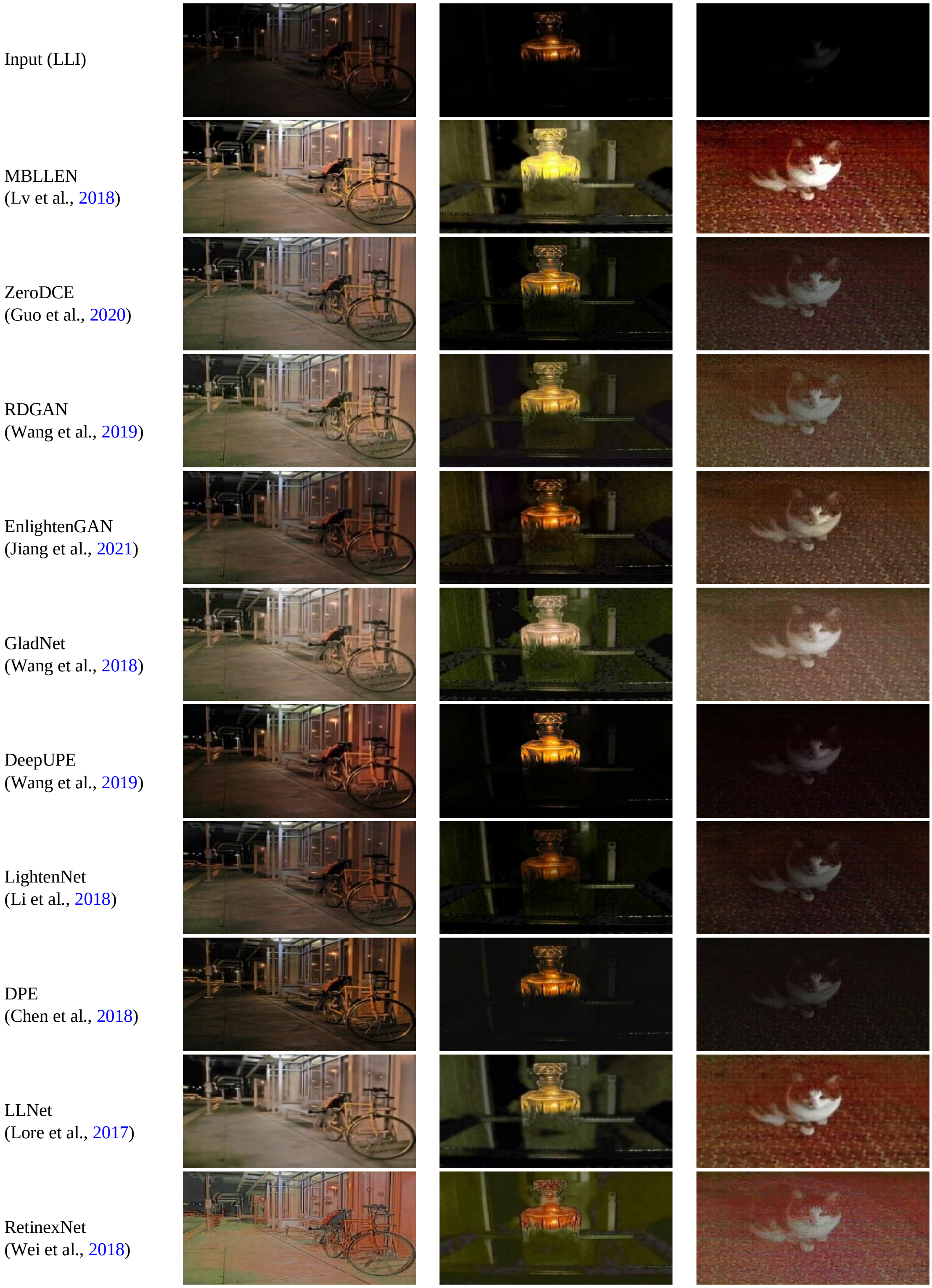
High-level computer vision tasks like object detection and classification usually suffer from a degraded performance when processing LLIs (Yang et al., 2020; VidalMata et al., 2020). In this experiment, we aim to verify whether enhancing LLI illumination and quality would improve the performance of the object detection task. To do so, we perform a comparative analysis using 4 object detection models: YOLOv3 (You Only Look Once version 3)[17] (Redmon and Farhadi, 2018), RetinaNet[18] (Lin et al., 2017), SSD (Single Shot MultiBox Detector)[19] (Wei et al., 2016), and Mask RCNN (Region based CNN)[20] (He et al., 2017). We apply the models on the entire original ExDark dataset as well as its enhanced versions produced by the 10 enhancement models considered in our previous experiment.

[17] https://github.com/ultralytics/yolov3

[18] https://github.com/fizyr/keras-retinanet

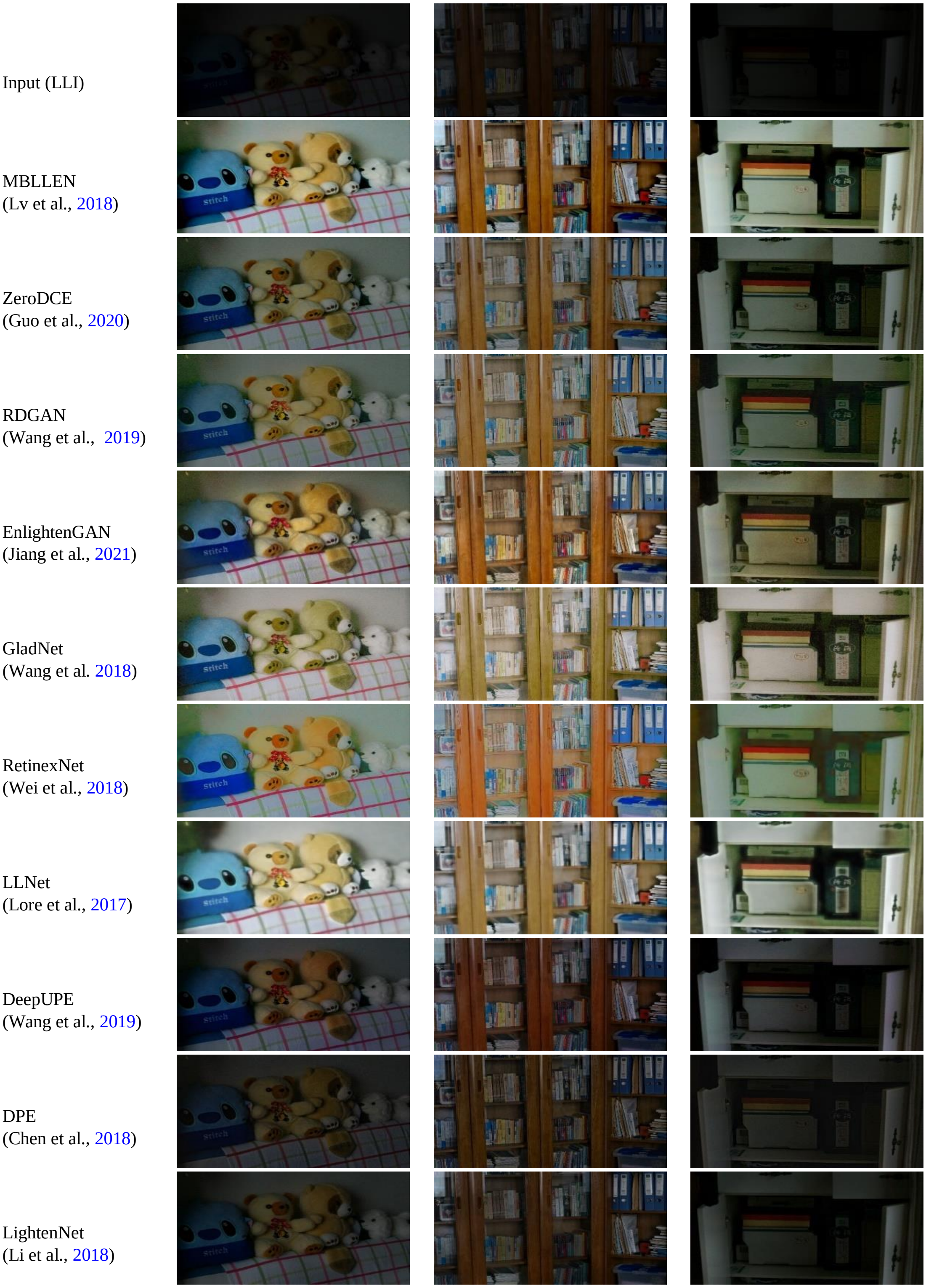
[19] https://github.com/pierluigiferrari/ssd_keras

[20] https://github.com/matterport/Mask_RCNN


Input (LLI)
MBLLEN (Lv et al., 2018)
ZeroDCE (Guo et al., 2020)
RDGAN (Wang et al., 2019)
EnlightenGAN (Jiang et al., 2021)
GladNet (Wang et al., 2018)
DeepUPE (Wang et al., 2019)
LightenNet (Li et al., 2018)
DPE (Chen et al., 2018)
LLNet (Lore et al., 2017)
RetinexNet (Wei et al., 2018)

(a) Sample 1 (b) Sample 2 (c) Sample 3

Fig. 5. Visual human comparison of enhanced LLIs from ExDark dataset, ordered from best to worst



(a) Sample 1 (b) Sample 2 (c) Sample 3

Fig. 6. Visual human comparison of enhanced LLIs from the LOL dataset, ordered from best to worst

6.1. Experimental Setup

We use the detection models' recommended weights, pre-trained on Microsoft COCO (Lin et al., 2014): a large-scale object detection, segmentation, and captioning dataset. This allows a generic evaluation, rather than training and fine-tuning the detection models using ExDark's LLIs or their enhanced counterparts, which would defeat the purpose of the experiment. After all, we aim to enhance LLIs to make them usable by existing detection models trained on large benchmark datasets of real world NLIs that are abundantly available. To do so, we leverage the ground truth bounding box annotations provided in the ExDark dataset to perform our experiments. We post-process the predictions provided by the detection models trained on COCO and fit them to ExDark. The COCO dataset consists of 80 different classes of objects, and ExDark consists only of 12 classes all of which are included in COCO. Few ExDark classes are more generic than their COCO counterparts, e.g., *couch* and *bench* classes in COCO are annotated as *chair* in ExDark, *wine glass* in COCO is annotated as *cup* in ExDark, and *truck* in COCO is annotated as *car* in ExDark. Hence, we match the classes from both datasets by converting COCO's predictions to the equivalent class annotations in ExDark and ignoring all the classes predicted by COCO that do not exist in ExDark, thus bounding the detections to ExDark's 12 classes. As for the comparison task, we utilize the mean Average Precision *(mAP*) metric commonly used to evaluate the performance of object detection and classification models in the literature (cf. Section 4.2.2).

6.2. Experimental Results

Table 6 presents the *mAP* results achieved by the 4 detection models, applied on the original ExDark dataset and its enhanced versions produced by the 10 enhancement models considered in our study. Results highlight the following observations:

- **The enhancement models produce consistent results:** They rank almost the same across three of the four considered detection models, except for YOLOv3 where RDGAN and GladNet are ranked as top third and fourth best models. As for the top-ranked and bottom-ranked models, they show consistency across the four detection models. Namely, MBLLEN and DeepUPE interchangeably rank at first and second top positions, while at the other end of the spectrum, RetinexNet and LLNet share the last two worst positions.
- **Minimal enhancement models produce good detection results:** DeepUPE, LightenNet, and DPE which perform minimal enhancement and show low illumination levels in Experiment 1 (cf. Section 5.2), are consistently ranked among the best models in terms of detection performance across all detections models except for DPE with YOLOv3 in which it ranks at the seventh position. This can be reasoned to their minimalistic enhancement, which keeps the enhanced images attached to their original LLIs, and thereby preserves their original features and semantics.
- **Object detection quality does not always correlate with visual enhancement quality**: One example is EnlightenGAN which ranks as the third worst model following detection performance using YOLOv3, and as the fifth model following other detection models. Yet, it consistently produces some of the best LLI enhancement results in Experiment 1. EnlightenGAN uses a reference-less self-feature preserving loss based on a pre-trained VGG16 model (Simonyan and Zisserman, 2015), which may not be able to effectively preserve the image features to itself, thus showing a degraded detection performance. On the opposite side of the spectrum, DeepUPE produces some of the best object detection results in this experiment, despite showing non-promising LLI enhancement results in Experiment 1. DeepUPE utilizes a pre-trained VGG16 encoder network to extract the image features before enhancing it. The powerful feature extraction capabilities of VGG16 might be a reason behind its good detection performance. Hence, an improved LLI visual enhancement quality does not seem to directly translate into improved detection and classification quality.
- **Few exceptions to the previous observations:** Results produced by MBLLEN and RetinexNet tend to contradict some of the previous observations. On the one hand, MBLLEN boasts some of the best enhancement quality levels on ExDark from Experiment 1, and consistently exhibits the second-best *mAP* levels across most detection models in the present experiment. The good *mAP* results can be attributed to MBLLEN's large-scale

training dataset: PASCAL VOC (Everingham et al., 2012) consisting of 16,925 images containing dynamic objects and classes similar to those in the ExDark dataset, thus probably allowing for a better preservation of the image visual contents and semantics. On the other hand, RetinexNet bears some of the worst enhancement quality levels and produces the worst object detection quality levels. This means that image enhancement quality is not completely disassociated from detection quality and can affect the object detection task.

Table 6. *mAP* results for the ExDark dataset, ranked from best (#1) to worst (#10) following each detection model (red color refers to the best score and green to the second best for every detection model)

(a) Results ranked following YOLOv3 (Redmon and Farhadi, 2018)

Approach	Rank	YOLOv3	RetinaNet	SSD	Mask RCNN
MBLLEN	1	**63.35**	**58.90**	**57.61**	**56.18**
DeepUPE	2	**63.24**	**59.14**	**57.74**	**56.64**
RDGAN	3	62.48	54.30	54.75	52.59
GladNet	4	62.37	54.09	54.89	52.40
LightenNet	5	62.30	57.77	56.51	54.66
ZeroDCE	6	62.20	54.76	55.26	51.80
Original LLIs	---	61.79	60.14	57.83	56.43
DPE	7	61.79	57.10	55.93	54.07
EnlightenGAN	8	61.61	55.47	55.60	53.75
LLNet	9	58.44	53.39	53.00	49.31
RetinexNet	10	51.08	40.19	43.18	34.69

(b) Results ranked following RetinaNet (Lin et al., 2017)

Approach	Rank	YOLOv3	RetinaNet	SSD	Mask RCNN
Original LLIs	---	61.79	60.14	57.83	56.43
DeepUPE	1	**63.24**	**59.14**	**57.74**	**56.64**
MBLLEN	2	**63.35**	**58.90**	**57.61**	**56.18**
LightenNet	3	62.30	57.77	56.51	54.66
DPE	4	61.79	57.10	55.93	54.07
EnlightenGAN	5	61.61	55.47	55.60	53.75
ZeroDCE	6	62.20	54.76	55.26	51.80
RDGAN	7	62.48	54.30	54.75	52.59
GladNet	8	62.37	54.09	54.89	52.40
LLNet	9	58.44	53.39	53.00	49.31
RetinexNet	10	51.08	40.19	43.18	34.69

(c) Results ranked following SSD (Wei et al., 2016)

Approach	Rank	YOLOv3	RetinaNet	SSD	Mask RCNN
Original LLIs	---	61.79	60.14	57.83	56.43
DeepUPE	1	**63.24**	**59.14**	**57.74**	**56.64**
MBLLEN	2	**63.35**	**58.90**	**57.61**	**56.18**
LightenNet	3	62.30	57.77	56.51	54.66
DPE	4	61.79	57.10	55.93	54.07
EnlightenGAN	5	61.61	55.47	55.60	53.75
ZeroDCE	6	62.20	54.76	55.26	51.80
GladNet	7	62.37	54.09	54.89	52.40
RDGAN	8	62.48	54.30	54.75	52.59
LLNet	9	58.44	53.39	53.00	49.31
RetinexNet	10	51.08	40.19	43.18	34.69

(d) Results ranked following Mask RCNN (He et al., 2017)

Approach	Rank	YOLOv3	RetinaNet	SSD	Mask RCNN
DeepUPE	1	**63.24**	**59.14**	**57.74**	**56.64**
Original LLIs	---	61.79	60.14	57.83	56.43
MBLLEN	2	**63.35**	**58.90**	**57.61**	**56.18**
LightenNet	3	62.30	57.77	56.51	54.66
DPE	4	61.79	57.10	55.93	54.07
EnlightenGAN	5	61.61	55.47	55.60	53.75
RDGAN	6	62.48	54.30	54.75	52.59
GladNet	7	62.37	54.09	54.89	52.40
ZeroDCE	8	62.20	54.76	55.26	51.80
LLNet	9	58.44	53.39	53.00	49.31
RetinexNet	10	51.08	40.19	43.18	34.69

6.3. Discussion

To summarize the above observations: i) most enhancement models produce consistent results and behave similarly across the object detection models evaluated in our study, ii) object detection quality does not always correlate with visual enhancement quality, where some good enhancement models may perform badly when used for object detection, and vice versa, and iii) a few exceptions to the previous observation show that image enhancement quality is not completely disassociated from object detection quality, and can improve the object detection task. Interestingly, most detection models (with exception of YOLOv3) tend to perform better on the original LLIs, compared with the enhanced images. While this seems counter-intuitive, yet a similar observation was made in a recent study in (VidalMata et al., 2020), where the authors summarize the results of the UG^2 Challenge workshop held at IEEE CVPR 2018[21], and which aims at assessing the influence of image restoration and enhancement techniques in improving the performance of classification models like VGG16 and VGG19 (Simonyan and Zisserman, 2015), InceptionV3 (Szegedy et al., 2016), and ResNet50 (He et al., 2016). Extensive experimentation on a new video

[21] 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18-22, 2018

benchmark dataset representing both ideal conditions and common aerial image artifacts, demonstrate that improving image quality does not necessarily lead to an improved classification performance, and may even degrade it in certain cases where images include extreme artifacts. To further clarify this point, we provide the precision-recall (*P-R*) curves for each of the 12 classes of the ExDark dataset in Appendix C. On the one hand, the curves in Fig. C.1-to-Fig. C.4 show that precision drops significantly with the increase of recall for both the original and enhanced images. In other words, *P-R* does not seem to improve with the enhancement task. On the other hand, average precision (*AP*) results per class for most detection models show better performance on the original LLIs compared with their enhanced counterparts. Note that *mAP* results for YOLOv3 show partial improvement with few enhancement models like MBLLEN and DeepUPE, especially when evaluated on certain classes like *Cat*, *Cup*, *Dog*, and *Motorbike* where objects are of small size and occur at multiple scales in which YOLOv3 is known to perform well. This improvement in detection quality, while being partial and local to a few models and object classes, suggests that LLI enhancement can help improve object detection performance if designed in a special way to highlight and preserve the features of interest to the object detection task.

Also, one aspect that might affect object detection quality is the level of noise added in the enhanced images. By comparing with the results of Experiment 1, we realize that MBLLEN produces some of the lowest noise levels compared with the other enhancement models (Table 4c) while producing some of best detection results in this experiment. ZeroDCE and RDGAN which are ranked among the best enhancement models by human testers in Experiment 1 (Fig. 4), produce some of the worst noise levels (Table 4c) and show a degraded detection performance with almost all detection models in this experiment. This suggests that a proper balancing between visual features and noise levels should be maintained to improve the detection task.

7. Experiment 3: Feature Analysis

In this experiment, we attempt to better understand the impact of LLI enhancement on high-level computer vision tasks by comparing the feature maps extracted from LLIs, NLIs, and their enhanced counterparts. To our knowledge, this is the first quantitative feature map evaluation study of its kind in the literature.

7.1. Experimental Setup

We present and discuss the results obtained by the SSD detection model (Wei et al., 2016) applied on images from the LOL dataset. We particularly utilize SSD512, a variant of SSD using 512 input image size pre-trained using the Microsoft COCO dataset (Lin et al., 2014). We chose this detection model for our analysis because of its convenient architecture (VGG16+6 extra sequential feature layers), comprising a sequence of stacked convolutional and pooling layers, and allowing to easily extract the feature maps of interest. In this experiment, we consider the feature maps from three sample layers of the detection model: i) a sequence of large maps from conv4_3 (64x64x512) – one of the layers belonging to the model's backbone, ii) a sequence of smaller maps from conv8_2 (16x16x512) – an intermediary layer, and iii) a sequence of minimal size maps from conv11_2 (1x1x256) – the model's last layer (Fig. 7).

We introduce two new metrics to compare the feature maps at a given layer of the DL model: i) Feature Map Matrix Similarity (*FMMS*), and ii) top N Active Feature Map Similarity (*topN-AFMS*). *FMMS* computes the cosine similarity measure between the feature maps of two (sets of) images at given layer of the DL model, highlighting overall image feature similarity (Equation (5) in Section 4.2.3). Here, we expect that the enhanced images share maximum feature similarity (producing maximum *FMMS* scores) with NLIs, compared with their LLI counterparts. In other words, we expect a high-quality enhanced image to share more similar features with a NLI, compared with a LLI. *topN-AFMS* compares the most active feature maps between two sets of pair-wise matching images, to help describe the behavior of the detection model and its response activity against the input images (Equation (7) in Section 4.2.3). The most active and responsive feature maps are those having the highest average activation at a certain layer of the DL model, while feature maps having zero average activations refer to dead or inactive maps. Identifying and comparing the most active feature maps gives insight into the features that might be most impactful on object detection and classification quality.

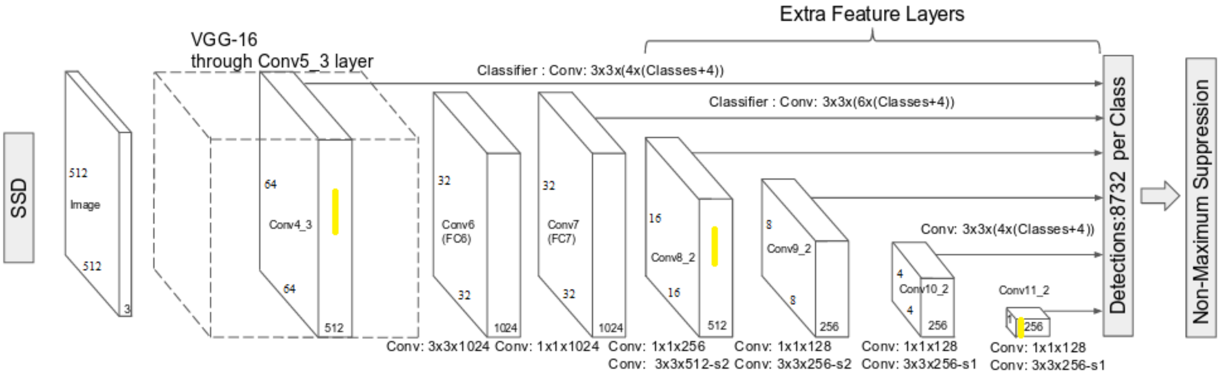
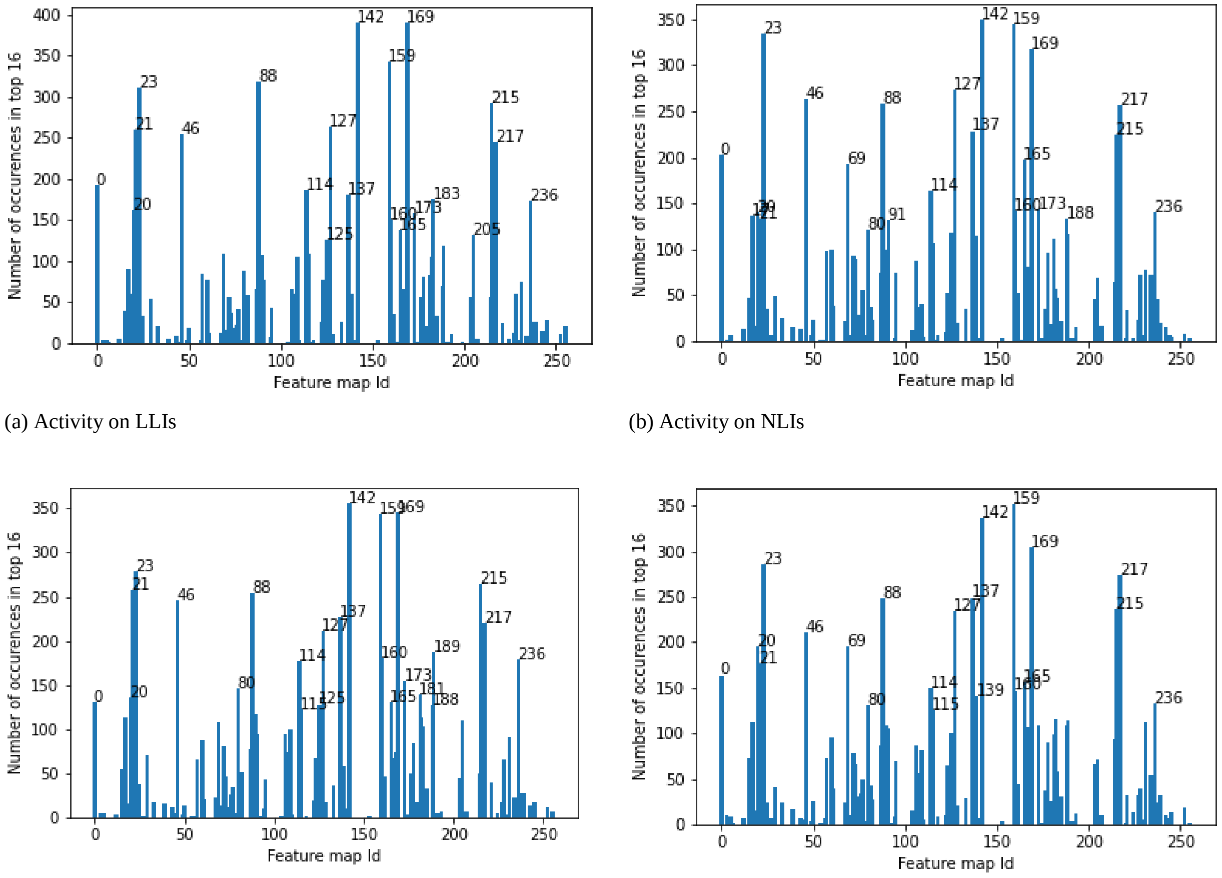


Fig. 7. SSD512 architecture: layers marked in yellow are used in the analysis (modified based on (Wei et al., 2016))



(a) Activity on LLIs

(b) Activity on NLIs

(c) Activity on enhanced images using DPE (Chen et al., 2018)

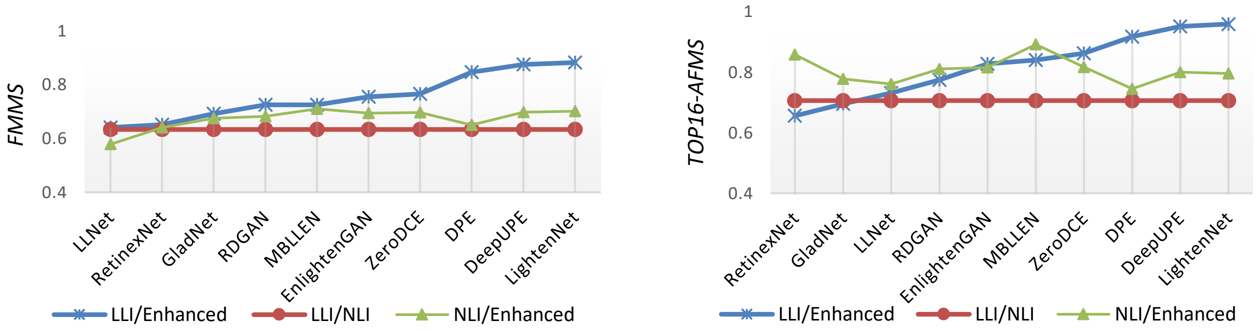
(d) Activity on enhanced images using MBLLEN (Lv et al., 2018)

Fig. 8. Visualizing top-16 active feature maps for layer conv11_2 for the LOL dataset. The labels on top of the bars refer to the feature map id within the layer

For instance, Fig. 8 shows the number of occurrences of the top-16 active feature maps extracted using SSD at layer conv11_2, considering all the images from the LOL dataset. One can see that feature maps 142 and 169 are the most active with about 400 occurrences among the LLIs (Fig. 8a), while feature maps 142 and 159 are the most active with about 350 occurrences among the top-16 active maps for NLIs (Fig. 8b). Note that images enhanced using the DPE model (which produced the worst *SSIM* and *PSNR* scores for LOL and minimal illumination levels in Experiment 1 – Table 5c, d) produce activity maps which are similar to those of LLIs, whereas images enhanced using MBLLEN (which produced the second best *NIQE* and the best *BRISQUE* scores – Table 5a, b – as well as the best subjective scores for LOL) produce activity maps which are similar to those of NLIs. Here, we expect that enhanced images share more similar active feature maps (producing higher *top16-AFMS* scores) with NLIs, compared with their LLI counterparts. In other words, we expect an enhanced image to preserve the most important (active) features that would be present in a NLI, compared with a LLI which tends to loosen certain image features.

7.2. Experimental Results

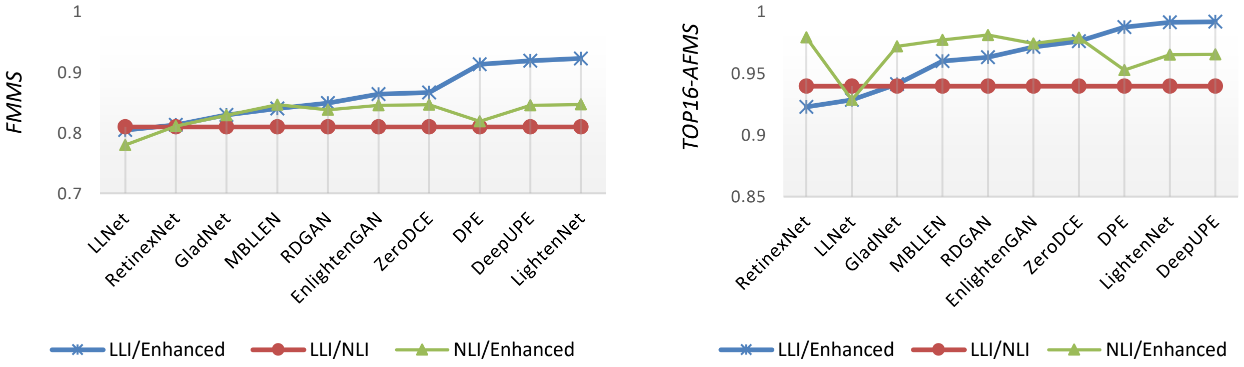
Fig. 9, Fig. 10, and Fig. 11 show the results of *FMMS* and *top16-AFMS* metrics respectively comparing LLIs/NLIs, LLIs/enhanced, and NLIs/enhanced images from the LOL dataset. Based on the obtained results, we highlight the following observations:



(a) *FMMS* results

(b) *top16-AFMS* results

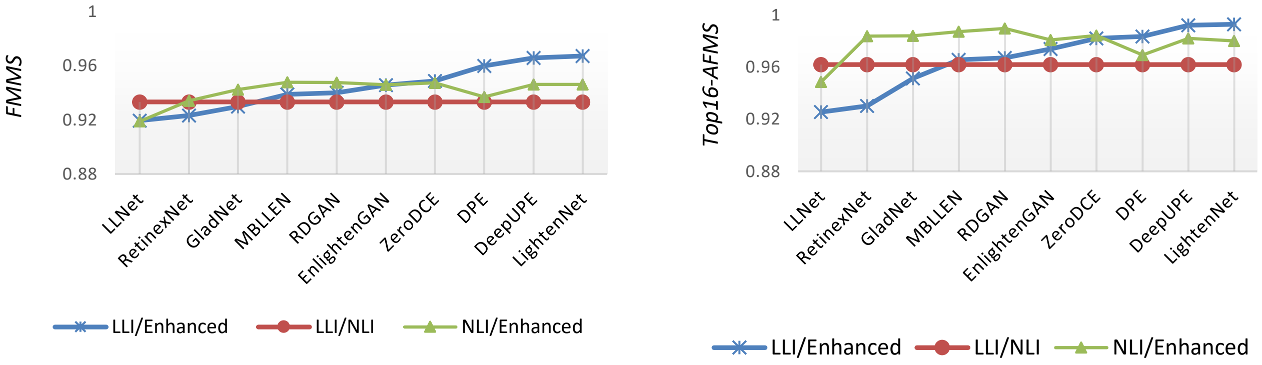
Fig. 9. Similarity measures for layer conv4_3 of SSD512 (Wei et al., 2016) applied on the LOL dataset, ordered following LLI/Enhanced



(a) *FMMS* results

(b) *top16-AFMS* results

Fig. 10. Similarity measures for layer conv8_2 of SSD512 (Wei et al., 2016) applied on the LOL dataset, ordered following LLI/Enhanced



(a) *FMMS* results

(b) *top16-AFMS* results

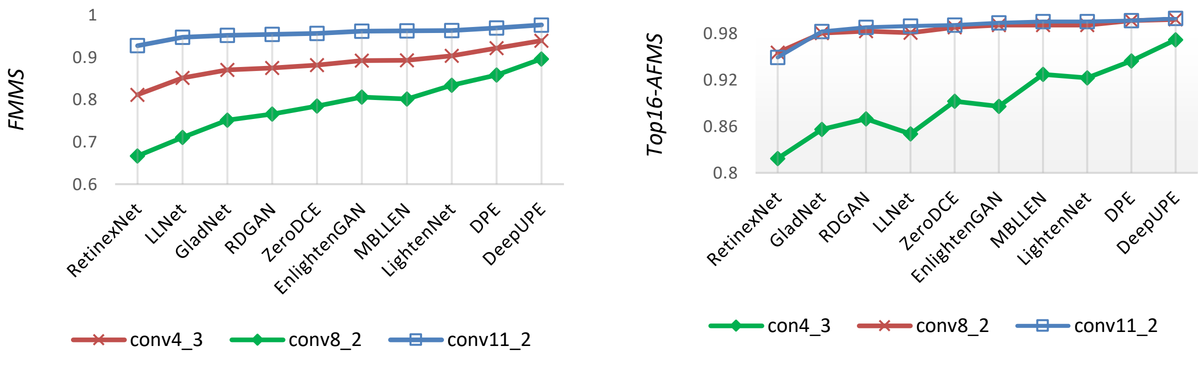
Fig. 11. Similarity measures for layer conv11_2 of SSD512 (Wei et al., 2016) applied on the LOL dataset, ordered following LLI/Enhanced

- ***Feature Map Matrix Similarity*** (*FMMS*) results for conv4_3 in Fig. 9a show that *FMMS*(LLI, enhanced) > *FMMS*(NLI, enhanced) for most enhancement models. This means that the enhanced images tend to share more features at conv4_3 with their LLI counterparts, compared with the corresponding NLIs, and thus remain attached to their initial LLIs. Almost the same pattern holds for conv8_2 in Fig. 10a (with the exception of MBLLEN). However, results at conv11_2, i.e., the deepest layer of SSD, show that most models produce maps which are closer to those of NLIs versus LLIs, except for DPE, DeepUPE, and LightenNet which remain largely attached to the initial LLIs. Results for the latter three models might be due to their minimal enhancement (Experiment 1 in Section 5.2) which makes them more faithful to the original LLIs. Additionally, LLNet shows the lowest *FMMS*(NLI, enhanced) and *FMMS*(LLI, enhanced) levels for all layers and more prominently for conv11_2 in Fig. 11a. This can be due to the over smoothing applied by LLNet on the enhanced images (Experiment 1) making them loose their fine details, especially in the deepest layers of the detection model (i.e., conv11_2) where the high-level features incorporated in the fine details are out of interest. Finally, MBLLEN shows some of the best results with *FMMS*(LLI, enhanced) approaching *FMMS*(LLI, NLI) and simultaneously producing approximately the highest *FMMS*(NLI, enhanced) scores compared with all other models and in all three layers. Other models do not share similar measures, for example LLNet produces very close *FMMS*(LLI, enhanced) and *FMMS*(LLI, NLI) scores in both conv4_3 and conv8_2 layers, and yet it shows the lowest *FMMS*(NLI, enhanced) score. Moreover DeepUPE has very close *FMMS* (NLI, enhanced) to that of MBLLEN, yet it shows much higher *FMMS*(LLI, enhanced).

- ***Top 16 Active Feature Map Similarity*** (*top16-AFMS*) results show that *top16-AFMS*(NLI, enhanced) > *top16-AFMS*(LLI, enhanced) in all layers and for most enhancement models except for DPE, DeepUPE, and LightenNet. This means that most enhancement models tend to activate the same feature maps in the detection model (e.g., SSD in this case) compared with NLIs, and succeed to diverge from the most active feature maps of the LLIs towards those of the NLIs. We also notice from Fig. 9b, 10b and 11b, that MBLLEN's *top16-AFMS*(LLI, enhanced) is approaching *top16-AFMS*(LLI, NLI) in all layers such that its *top16-AFMS*(NLI, enhanced) produces high scores compared with all other models. This shows that MBLLEN's enhanced images share very similar active features with their NLI counterparts, compared with their LLIs. MBLLEN's *FMMS* and *top16-AFMS* scores seem to corroborate the results of Experiment 1 (cf. Section 5.1 and 5.2) where the model produces one of the best enhancement quality results compared with the other models on the LOL dataset. To sum up, most enhancement models fall short of simultaneously producing high *FMMS*(NLI, enhanced) and high *top16-AFMS*(NLI, enhanced) along with *FMMS*(LLI, enhanced) ≈ *FMMS*(LLI, NLI) and *top16-AFMS*(LLI, enhanced) ≈ *top16-AFMS*(LLI, NLI), which suggest that the enhanced images remain attached to the original LLIs and do not diverge towards the actual NLIs.

We further apply our feature analysis to 500 sample images from the ExDark dataset. Here, we only compute *FMMS* and *top16-AFMS* for LLI/enhanced images since the dataset does not include NLIs. Based on the results in Fig. 12, we highlight the following observations:

- The behavior of SSD seems to be different between LOL and ExDark datasets. For instance, results in Fig. 12b show that the *top16-AFMS*(LLI, enhanced) for LLNet at conv11_2 is higher than those of RetinexNet, GladNet and RDGAN when applied on ExDark, while LLNet produces the lowest scores when applied on LOL (Fig. 11b). Similar fluctuations occur with the other enhancement models, producing different activity responses when applied on the quasi-synthetic LOL dataset, compared with the real-world ExDark.



(a) *FMMS* similarity results

(b) *top16-AFMS* similarity results

Fig. 12. Similarity measures between LLIs and enhanced images for SSD512 (Wei et al., 2016) applied on the ExDark subset, ordered following conv11_2 results

- Most enhancement models (except for RetinexNet) produce high *FMMS*(LLI, enhanced) and *top16-AFMS*(LLI, enhanced) scores and show minimal variations compared with the results produced with the LOL dataset (Fig. 11a, b). This means that the enhanced images produced by most models remain mostly faithful to their original LLIs and do not diverge enough from the LLIs to promote better features that can be more useful for the object detection task.
- DeepUPE, DPE and LightenNet show the highest *FMMS*(LLI, enhanced) and *top16-AFMS*(LLI, enhanced). This indicates that enhanced images produced by these models remain largely attached to the original LLIs and are thereby consistent with their minimal enhancement reflected in Experiment 1 (cf. Section 5.2). It may also explain their good object detection performance in Experiment 2 (cf. Section 6.2) since they tend to preserve the original LLI's semantic features.

7.3. Occlusion Experiment

To better interpret and understand the effect of enhancement on the preservation of image features and what may be lost during enhancement, we utilize the occlusion experiment proposed in (Zeiler and Fergus, 2014) where: i) a black square is used to mask particular sections of the image, ii) the black square is slid over all the possible sections of the image, allowing to iii) perform object detection for every mask, producing a heatmap highlighting the object detection confidence scores (in case of a detection, and zero otherwise). The occlusion experiment is performed on images containing a single object each, so that the detection model focuses solely on them. Its rationale is two-fold: i) if an image contains many regions which may cause a misdetection if occluded, then the image is assumed to hold weak features allowing to easily misdetect its object, and ii) if an image contains no specific region that might cause a misdetection if occluded – given all the masks slid over the entire image, then the image is assumed to hold strong features that allow to correctly detect its object.

In the following, we present both quantitative and qualitative evaluations of the occlusion experiment applied on enhanced images produced by the 10 DL-based LLI enhancement models used in the previous experiments.

7.3.1. Quantitative Evaluation

We perform the occlusion experiment on 100 sample images from the ExDark dataset, considering only images containing single objects. All the images are resized equally to 600×700, and the same size of the moving black box is used with all of them. We consider the original LLIs and their enhanced counterparts produced by the 10 enhancement models considered in our study, and process each of them through the 4 object detection models used in Experiment 2 (cf. Section 6). A threshold confidence score of 0.15 is used to limit the number of detections produced by all models. We make use of the Occlusion based Average Misdetection Regions (*OAMR*) metric (cf. Section 4.2.3), which highlights the ability of an enhancement model to include stronger features in the enhanced images by producing lower scores (i.e., lower number of regions contributing to misdetections) compared with the original LLIs.

Table 7. OAMR ↓ results for 100 images from ExDark dataset, ranked from best (#1) to worst (#10) following each detection model (red color refers to the best score and green to the second best for every detection model)

(a) Results ranked following YOLOv3 (Redmon and Farhadi, 2018)

Approach	**Rank**	**YOLOv3**	**RetinaNet**	**SSD**	**Mask RCNN**
Original LLIs	---	5.70	6.31	5.81	8.17
DeepUPE	1	**5.82**	**6.53**	**5.38**	**7.51**
MBLLEN	2	**6.42**	7.31	**5.76**	9.93
DPE	3	6.61	**5.58**	5.78	11.16
LightenNet	4	6.75	8.34	6.37	9.48
EnlightenGAN	5	7.40	9.56	6.68	9.66
GladNet	6	7.49	9.51	8.05	**9.40**
RDGAN	7	7.70	9.06	7.73	10.74
ZeroDCE	8	7.99	9.69	8.28	11.02
LLNet	9	8.83	11.23	7.46	11.89
RetinexNet	10	18.57	22.35	18.25	26.04

(b) Results ranked following RetinaNet (Lin et al., 2017)

Approach	**Rank**	**YOLOv3**	**RetinaNet**	**SSD**	**Mask RCNN**
DPE	1	6.61	**5.58**	5.78	11.16
Original LLIs	---	5.70	6.31	5.81	8.17
DeepUPE	2	**5.82**	**6.53**	**5.38**	**7.51**
MBLLEN	3	**6.42**	7.31	**5.76**	9.93
LightenNet	4	6.75	8.34	6.37	9.48
RDGAN	5	7.70	9.06	7.73	10.74
GladNet	6	7.49	9.51	8.05	**9.40**
EnlightenGAN	7	7.40	9.56	6.68	9.66
ZeroDCE	8	7.99	9.69	8.28	11.02
LLNet	9	8.83	11.23	7.46	11.89
RetinexNet	10	18.57	22.35	18.25	26.04

(c) Results ranked following SSD (Wei et al., 2016)

Approach	**Rank**	**YOLOv3**	**RetinaNet**	**SSD**	**Mask RCNN**
DeepUPE	1	**5.82**	**6.53**	**5.38**	**7.51**
MBLLEN	2	**6.42**	7.31	**5.76**	9.93
DPE	3	6.61	**5.58**	5.78	11.16
Original LLIs	---	5.70	6.31	5.81	8.17
LightenNet	4	6.75	8.34	6.37	9.48
EnlightenGAN	5	7.40	9.56	6.68	9.66
LLNet	6	8.83	11.23	7.46	11.89
RDGAN	7	7.70	9.06	7.73	10.74
GladNet	8	7.49	9.51	8.05	**9.40**
ZeroDCE	9	7.99	9.69	8.28	11.02
RetinexNet	10	18.57	22.35	18.25	26.04

(d) Results ranked following Mask RCNN (He et al., 2017)

Approach	**Rank**	**YOLOv3**	**RetinaNet**	**SSD**	**Mask RCNN**
DeepUPE	1	**5.82**	**6.53**	**5.38**	**7.51**
Original LLIs	---	5.70	6.31	5.81	8.17
GladNet	2	7.49	9.51	8.05	**9.40**
LightenNet	3	6.75	8.34	6.37	9.48
EnlightenGAN	4	7.40	9.56	6.68	9.66
MBLLEN	5	**6.42**	7.31	**5.76**	9.93
RDGAN	6	7.70	9.06	7.73	10.74
ZeroDCE	7	7.99	9.69	8.28	11.02
DPE	8	6.61	**5.58**	5.78	11.16
LLNet	9	8.83	11.23	7.46	11.89
RetinexNet	10	18.57	22.35	18.25	26.04

Results are reported in Table 7 and highlight the following observations:

- DeepUPE, MBLLEN, and DPE produce some of the best (low) *OAMR* results, which is consistent with their high *mAP* scores obtained in Experiment 2 (cf. Section 6.2), despite DeepUPE and DPE's minimal enhancement quality in Experiment 1 (cf. Section 5.1 and 5.2). Note that EnlightenGAN, which ranks among the top models in terms of enhancement quality following Experiment 1, does not show remarkable *OAMR* results. This corroborates with our observations from Experiment 2, where a good enhancement quality does not necessarily translate into better feature preservation and improved object detection quality.

- None of the remaining enhancement models (with the exception of DeepUPE, MBLLEN, and DPE) produce an *OAMR* score lower than that of the original LLIs, indicating that the models are adding more regions which contribute to misdetections, and are thereby losing significant object features upon image enhancement.
- MBLLEN produces one of the best (lowest) average *OAMR* scores, reflecting good feature preservation in the enhanced images. This seems consistent with its top enhancement quality achieved in Experiment 1. On the other end of the spectrum, RetinexNet shows the worst (highest) average *OAMR* scores, which is consistent with its bad enhancement quality achieved in Experiment 1. This seems to indicate that visual enhancement quality and feature preservation performance might not be completely unrelated, and that good visual enhancement balanced with proper feature handling could strengthen the object features upon image enhancement.

7.3.2. Qualitative Evaluation

In addition to the quantitative evaluation, and in order to shed further light on the results of the occlusion experiment, we qualitatively evaluate and discuss two typical cases using the YOLOv3 detection model: i) successful detection in both the LLI and the enhanced image, and ii) successful detection in the LLI and misdetection in the enhanced image. The cases where we have a misdetection in the LLI are not beneficial for this experiment since they do not reflect any information about the initial LLI features**.**

***Case 1** – **Successful detection in the LLI and the enhanced image**: Fig. 13 shows the occlusion heatmaps obtained on a sample LLI and its enhanced counterparts.

Results in Fig. 13 highlight the following observations:

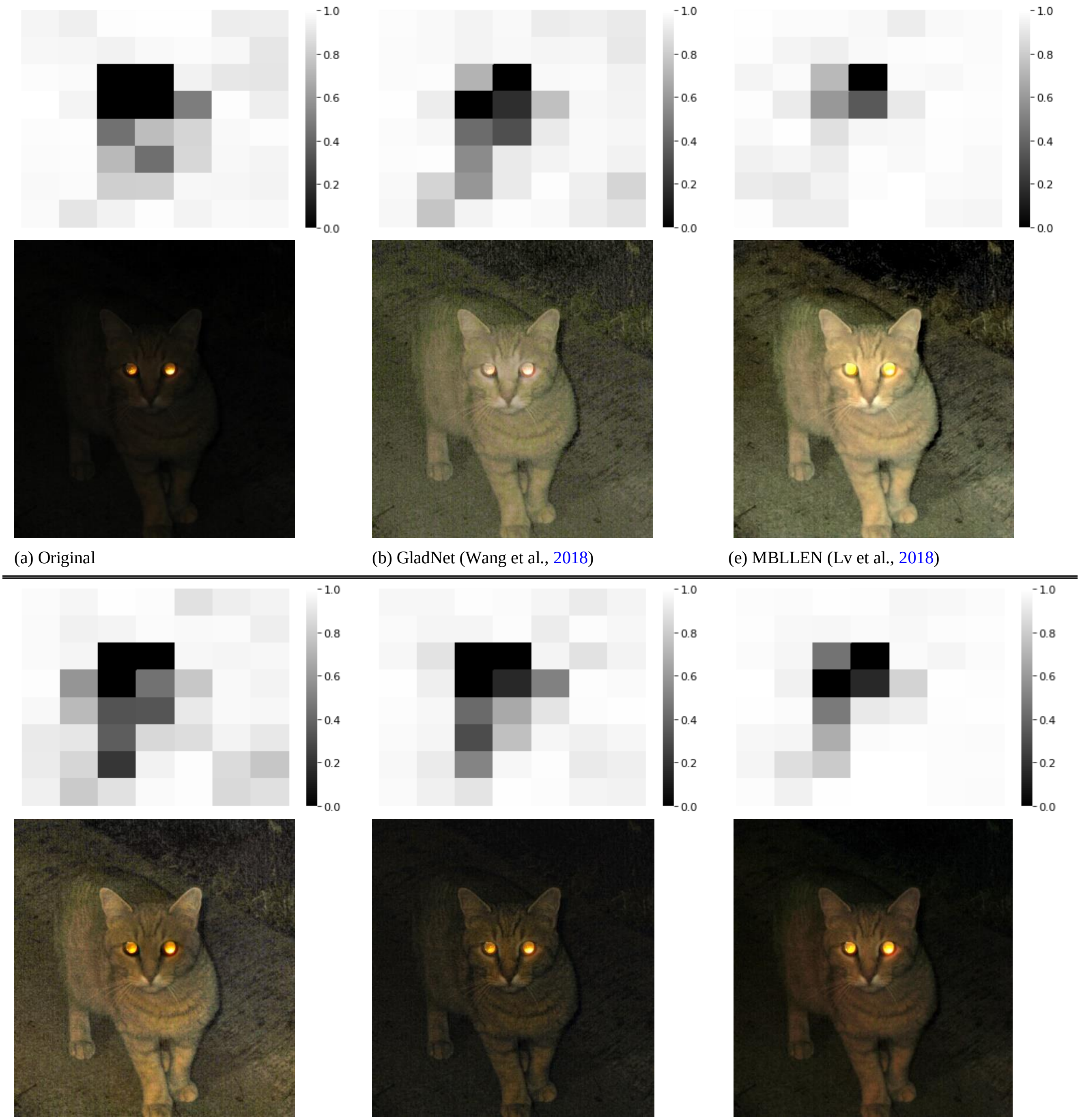
- The heatmap of the original LLI shows 4 zero-confidence score (dark) regions concentrated around the face of the *cat* object, which seem to contribute to its misdetection. In contrast, the heatmaps of the enhanced images show a lesser number of zero-score (dark) regions contributing to the misdetection of the *cat* object. This means that the enhancement models seem to integrate better features into the enhanced images, allowing to improve their detection confidence scores.

- MBLLEN shows the best features with only one region resulting in a misdetection. In other words, regions which were initially responsible for misdetecting the *cat* object in the original image are no longer causing a misdetection after MBLLEN' s enhancement.

- Although EnlightenGAN shows one of the best visual quality results among all enhancement models in Experiment 1 (Section 5.1), yet it produces 3 zero-confidence (dark) regions resulting in misdetections (Fig. 13d). In contrast, DeepUPE which shows a minimal enhancement quality in Experiment 1 produces only 2 regions resulting in misdetections (Fig. 13f). Similarly, DPE shows low illumination while producing only 3 misdetection regions (Fig. 13e), identically to EnlightenGAN which seemingly shows better illumination and enhancement quality (Fig. 13d). The latter observations show that a good visual enhancement quality does not necessarily translate into better object detection features. This corroborates the results from Experiments 1 and 2, where EnlightenGAN on the one hand, and DeepUPE and DPE on the other hand, respectively show high/low visual enhancement quality versus low/high object detection performance.

Case 2 – Successful detection in the LLI and misdetection in the enhanced image. Fig. 14 shows the occlusion heatmaps obtained on another sample LLI from the ExDark dataset, and its enhanced counterparts. Results highlight the following observations:

- Although object *cat* is successfully detected in the original LLI, yet it contains 11 regions contributing to a misdetection. This shows that the LLI initially holds weak features that poorly contribute to the object detection task.



(a) Original

(b) GladNet (Wang et al., 2018)

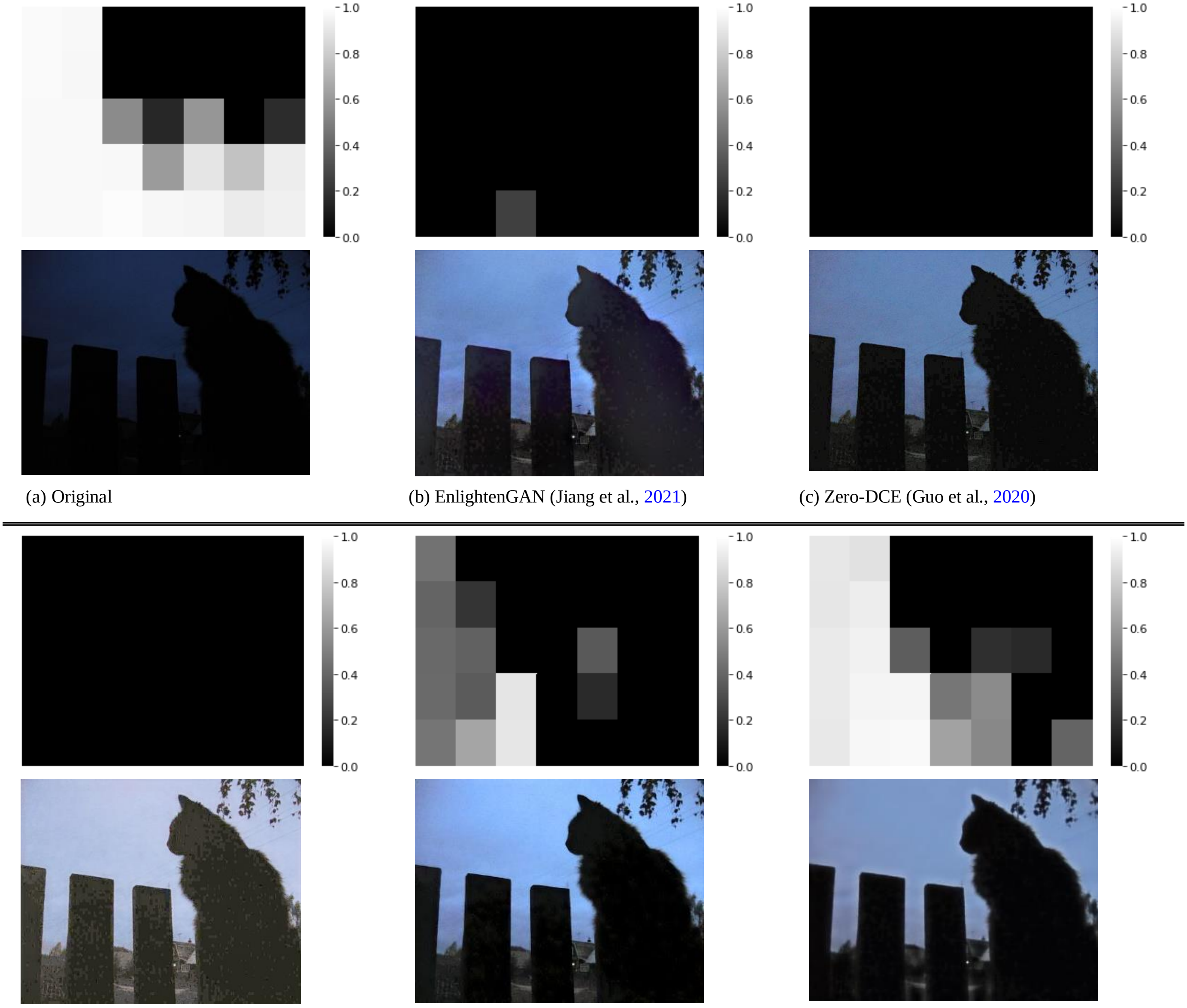
(e) MBLLEN (Lv et al., 2018)

(d) EnlightenGAN (Jiang et al., 2021)

(e) DPE (Chen et al., 2018)

(f) DeepUPE (Wang et al., 2019)

Fig. 13. Occlusion experiment on a sample LLI from ExDark and its enhanced counterparts: case 1



(a) Original

(b) EnlightenGAN (Jiang et al., 2021)

(c) Zero-DCE (Guo et al., 2020)

(d) GladNet (Wang et al., 2018)

(e) MBLLEN (Lv et al., 2018)

(f) LLNet (Lore et al., 2017)

Fig. 14. Occlusion experiment on a sample LLI from ExDark and its enhanced counterparts: case 2

- The enhanced images produced by EnlightenGAN, ZeroDCE, and GladNet, result in a complete misdetection of object *cat*, showing that the enhancement models have loosened the few features that were contributing to the object detection task in the original LLI.
- The enhanced image produced by MBLLEN allows a successful object detection. However, it includes 22 regions contributing to a misdetection, which is double the number of misdetection regions present in the original LLI (=11). This shows that MBLLEN loosened some of the features while preserving others that were most important to the detection task.

- LLNet allows a successful object detection and shows better feature preservation compared with its counterparts. This can be due to the minimal noise incorporated by LLNet in comparison with its counterparts (Experiment 1, Table 4c). This shows that the amplification or integration of noise into the enhanced image seems to loosen the features that are useful for object detection.

7.4. Discussion

To sum up, we review and discuss the results of our feature analysis experiment.

First, the enhanced images produced by most enhancement models tend to activate the same detection model feature maps compared with LLIs. In other words, enhanced images produced by most models tend to share more features with their LLI counterparts, compared with the corresponding NLIs. They fail to diverge from the features of the LLIs towards those of the NLIs and remain attached to their initial LLIs.

Second, results of the occlusion experiment show that successful object detection in enhanced images seems to be related to the number of (mis)detection regions in the occlusion heatmap, which in turn highlights the number of (loosened and) preserved features in the resulting enhanced image, compared with its original LLI. *OAMR* results show that most of the enhancement models tend to produce enhanced images with more regions contributing to misdetections and thus showing weakly embodied semantic features.

Third, an important aspect to be considered here is the level of noise added in the enhanced images. Referring to the results of Experiments 1 and 2, we realize that MBLLEN produces some of the lowest noise levels (cf. Section 5.1) and some of the best *mAP* results (cf. Section 6.2) compared with the other enhancement models, and accordingly produces good *OAMR* scores in this experiment. On the other side of the spectrum, ZeroDCE produces the highest noise level amongst the enhancement models (cf. Section 5.1) with uncompetitive *mAP* results (cf. Section 6.2), and accordingly produces some of the worst *OAMR* scores. This suggests that preserving the image features that are useful for object detection, coupled with a reduction in noise levels, can help improve detection performance.

8. Recap and Directions

8.1. Recap of challenges

We recap the challenges facing DL-based LLI enhancement models based on our literature review from Section 3:

- Reliance on supervised learning where paired LLIs/NLIs are needed to train the models. Collecting large datasets of real-world LLIs and their corresponding daytime counterparts for the same scenes is difficult and challenging. Many techniques utilize synthetic LLIs produced from NLIs using light correction and noise induction techniques, yet synthetic LLIs do not accurately represent real world low-light conditions.
- Struggling with low-quality, noisy, and extremely dark images, hence the need for a proper understanding and modeling of the quality and noise elements in an image when conducting image enhancement.
- Performing enhancement without taking into consideration high-level computer vision tasks like object detection and classification, where high-level image features might be distorted or lost during the enhancement task, thus leading to reduced or non-improving end-to-end performance.
- Difficulty in comparing different LLI models due to the lack of standard metrics and datasets for comparison and evaluation.

8.2. Recap of Empirical Results

We also summarize our observations and findings based on our empirical evaluations from Sections 5-7:

From *Experiment 1 - Visual and Perceptual Quality*:

- Most of the LLI enhancement models evaluated in our study still fall short of producing properly illuminated enhanced images with good visual quality. They fail to strike a good balance between image illumination level, noise level, exposure level, and color deviation. Some models successfully improve one aspect while ignoring others and tend to incorporate significant noise into the enhanced images, thus distorting their quality.

- Results for the IQA (Image Quality Assessment) objective metrics used in this study do not closely match human evaluation ratings. The metrics also fail to produce consistent rankings among themselves.

From Experiment 2 – Detection and Classification quality:

- Improving LLI visual quality does not necessarily boost object detection and classification quality. Many models evaluated in our study tend to produce enhanced images which deteriorate object detection performance rather than improving it. This can be attributed to the fact that most existing enhancement models were developed as standalone solutions, and were not designed to be embedded as a pre-processing step for high-level computer vision tasks like object detection.
- The level of noise added in the enhanced images seems to affect detection quality. By comparing with the results of Experiment 1, we realize that many models producing low noise levels tend to produce some of the best detection results in Experiment 2. This suggests that a proper balancing between noise level and visual features could improve the detection task.

From Experiment 3 – Feature Analysis:

- Enhanced images produced by most models tend to share more features with their LLI counterparts, compared with the corresponding NLIs. They fail to diverge from the features of the LLIs towards those of the NLIs, and remain attached to their initial LLIs. This contributes to a drop in detection performance, which is usually further exacerbated by the added artifacts and noise resulting from the enhancement process.
- Most enhancement models tend to produce enhanced images with more regions contributing to misdetections and thus show weakly embodied semantic features. The enhancement task should consider enriching enhanced images with strong features that make detection models more robust and confident in their predictions.

8.3. Potential Directions

Based on our literature review and empirical observations, we highlight the following potential directions:

- There is a need to design more accurate IQA objective metrics that simultaneously quantify illumination and noise levels and behave in accordance with the human visual perception of image quality.
- There is a need to produce LLI enhancement models that can be used as a pre-processing step for other high-level computer vision task such as object detection and classification. In a recent study in (Al Sobbahi and Tekli, 2022), the authors propose to integrate homomorphic filtering within a Deep Learning (DL) solution, performing image-to-frequency filter learning designed for seamless integration into state-of-the-art DL image classification solutions. Results are promising and show improved enhancement and classification quality. In this context, LLI enhancement should be formulated while considering the preservation of the semantic features necessary for the high-level task at hand. Training and fine-tuning the object detectors on the enhanced datasets directly, using for instance a subset of COCO on which LLI enhancement is applied (even if COCO images have good illuminations) might help the detector to learn how the processed images are represented, and thus potentially improve detection quality.
- The preservation of semantic features should consider decoupling the LLI from the enhanced image such that it does not diverge beyond the actual similarity between the NLI and LLI, while maintaining at the same time high similarity with the NLI. While most supervised learning models tend to use a perceptual loss between the enhanced image and the NLI, they should also consider limiting the loss between the LLI and the enhanced image to that between the LLI and NLI.
- Solutions targeting different kinds of ill-lighting conditions (e.g., low-light, back-lit, over-exposed, front-lit, and combinations of them) need to be reviewed and compared in a self-contained study. This is specifically relevant in the field of medical imaging, where the restricting anatomy of the human body and the technical

limitations of the recording equipment might result in different forms of insufficient illumination, which complicate clinical examination and analysis (Ma et al., 2021; Gomez et al., 2019). While the ExDark dataset includes images with varying exposure levels which we use to evaluate the empirical models considered in our present study, yet conducting an in-depth evaluation of the impact of different deep learning models and the challenges of different kinds of ill-lighting conditions and different kinds of low-light images remains a major future direction.

- The noise factor and de-noising techniques need to be given special attention when designing new LLI enhancement models, especially that noise seems to consistently affect visual enhancement quality as well as object detection quality. Various DL-based solutions have been recently proposed to perform denoising, e.g., (Tian et al., 2020; Su et al., 2019; Guan et al., 2019), and dehazing, e.g., (Sheng et al., 2022; Xu and Wei, 2022, Ting et al., 2022). The authors in (Li et al, 2020) put forward a dedicated DL solution for underwater image enhancement. They synthesize underwater image degradation datasets considering different water types and degradation levels, and then train a dedicated convolutional neural network (CNN) on each type, and extend it to perform frame-by-frame video enhancement. In this context, designing and integrating different DL-based LLI enhancement solutions to perform denoising, dehazing, and underwater enhancement, and comparing the different model types against each other, is a major future direction.
- One of the best enhancement models in our empirical evaluation: EnlightenGAN (Jiang et al., 2021), follows the unsupervised learning paradigm, and thus highlights the potential of unsupervised LLI enhancement techniques. This would eliminate the need for paired training images, and would allow the use of real-world datasets which are increasingly available, rather than relying on synthetic datasets which are scarce and fail to mimic real LLIs.
- Another promising direction is presented by ZeroDCE (Guo et al., 2020), which entirely reformulates the LLI enhancement task to learn a mapping between LLIs and estimated light curves, thus releasing the need of paired and unpaired training data. The model achieved good object detection quality compared with many other DL enhancement models and was qualitatively favored by human testers as it sufficiently boosted image illumination albeit adding more noise. Such an approach could be revolutionary if properly extended or fine-tuned to maintain a good balance between illumination level, noise level, and semantic feature preservation.

9. Conclusion

In this study, we gave an overview of current DL-based LLI enhancement models, which we organized in 5 main categories: encoder-decoder and CNN based, Retinex based, Fusion based, GAN based, and more recent Zero Reference based models. Then, we described the experimental evaluation and results comparing 10 of the most recent DL-based LLI enhancement models. We conducted three main experiments evaluating: i) visual and perceptual quality, where LLI enhancement models were evaluated as standalone applications, ii) detection and classification quality, achieved by 4 different object detection models applied on LLIs and their enhanced counterparts, where LLI enhancement models were embedded as a pre-processing step in the overall pipeline, and iii) feature analysis, considering the effect of LLI enhancement on the resulting image features and its impact on object detection performance. We finally summarized our empirical observations and highlighted various potential research directions. We hope that the unified presentation of DL-based LLI enhancement in this paper will contribute to strengthen further research on the subject.

Acknowledgments

This study is partly funded by the National Council for Scientific Research - Lebanon (CNRS-L), and by the Lebanese American University (LAU). We would like to thank the testers (including LAU master's and senior engineering students) who volunteered to participate in the experimental evaluation.

References

Abdullah-Al-Wadud, M., Kabir, M. H., Dewan, M. A., Chae, O., 2007. A Dynamic Histogram Equalization for Image Contrast Enhancement. IEEE Trans. Consum. Electron. 53, 593-600.

Abu-Khzam, F. N., Daudjee, K., Mouawad, A. E., Nishimura, N., 2015. On Scalable Parallel Recursive Backtracking. J. Parallel Distrib. Comput. 84, 65-75.

Abu-Khzam, F., Li, S., Markarian, C., Heide, F., Podlipyan, P., 2019. Efficient Parallel Algorithms for Parameterized Problems. Theor. Comput. Sci. 786, 2-12.

Abu-Khzam, F. N., Markarian, C., Heide, F. M., Schubert, M., 2018. Approximation and Heuristic Algorithms for Computing Backbones in Asymmetric Ad-hoc Networks. Theory Comput. Syst. 62, 1673-1689.

Agaian, S. S., Silver, B., Panetta, K. A., 2007. Transform Coefficient Histogram-Based Image Enhancement Algorithms Using Contrast Entropy. IEEE Trans. Image Process., 16, 741-758.

Agustsson, E., Timofte, R., 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1122-1131.

Al Sobbahi R. and Tekli J., 2022. Low-Light Homomorphic Filtering Network for Integrating Image Enhancement and Classification, Signal Processing: Image Communication (SPIC), https://doi.org/10.1016/j.image.2021.116527.

Amigó, J. M., Kocarev, L., Tomovski, I., 2007. Discrete Entropy. Physica D: Nonlinear Phenomena 228, 77-85.

Arbelaez, P., Maire, M., Fowlkes, C., Malik, J., 2011. Contour Detection and Hierarchical Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 33, 898–916.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein Generative Adversarial Networks. In: International Conference on Machine Learning (ICML). pp. 214-223.

Bileschi, S. M., 2006. Streetscenes: Towards Scene Understanding in Still Images. Massachusetts Inst. of Tech. Cambridge, Tech. Rep.

Blau, Y., Michaeli, T., 2018. The Perception-Distortion Tradeoff. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6228-6237.

Bychkovsky, V., Paris, S., Chan, E., Durand, F., 2011. Learning Photographic Global Tonal Adjustment with a Database of Input / Output Image Pairs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 97-104.

Cai, J., Gu, S., Zhang, L., 2018. Learning a Deep Single Image Contrast Enhancer from Multi-Exposure Images. IEEE Trans. Image Process. 27, 2049–2062.

Chen, C., Chen, Q., Xu, J., Koltun, V., 2018. Learning to See in the Dark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3291-3300.

Chen, Y.-S., Wang, Y.-C., Kao, M.-H., Chuang, Y.-Y., 2018. Deep Photo Enhancer: Unpaired Learning for Image Enhancement from Photographs with GANs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6306-6314.

Chen, Z., Abidi, B. R., Page, D. L., Abidi, M. A., 2006. Gray-level grouping (GLG): An Automatic Method for Optimized Image Contrast Enhancement-part I: The Basic Method. IEEE Trans. Image Process. 15, 2290-2302.

Cheng, Y., Yan, J., Wangy, Z., 2019. Enhancement of Weakly Illuminated Images by Deep Fusion Networks. In: IEEE International Conference on Image Processing (ICIP). pp. 924-928.

Dabov, K., Foi, A., Egiazarian, K., 2006. Image Denoising with Block Matching and 3D Filtering. In: Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning. pp. 354-365.

Dang-Nguyen, D., Pasquini, C., Conotter, V., Boato, G., 2015. RAISE: A Raw Images Dataset for Digital Image Forensics. In: ACM Multimedia Systems Conference (MMSys). pp. 219-224.

Deng, L., 2014. A Tutorial Survey of Architectures, Algorithms, and Applications for Deep Learning. APSIPA Trans. Signal Info. Process. 3.

Dong, C., Loy, C., Tang, X., 2016. Accelerating the Super-Resolution Convolutional Neural Network. In: European Conference on Computer Vision (ECCV). pp. 391-407.

Ebrahimi, D., Sharafeddine, S., Ho, P., Assi, C., 2021. Autonomous UAV Trajectory for Localizing Ground Objects: A Reinforcement Learning Approach. IEEE Trans. Mobile Comput. 20, 1312-1324.

Everingham, M., Gool, L. V., Williams, C. K., Winn, J., Zisserman, A., 2012. The Pascal Visual Object Classes (VOC) Challenge. Int. J. Comput. Vis. 88, 303–338.

Fu, X., Zeng, D., Huang, Y., Zhang, X.-P., Ding, X., 2016. A Weighted Variational Model for Simultaneous Reflectance and Illumination Estimation. In: IEEE Conference on Computer Vision & Pattern Recognition (CVPR). pp. 2782–2790.

Gharbi, M., Chen, J., Barron, J. T., Hasinoff, S. W., Durand, F., 2017. Deep Bilateral Learning for Real-time Image Enhancement. ACM Trans. Graph. 36, 118:1-118:12.

Gómez P., Semmler M., Schützenberger A., Bohr C., Döllinger M., 2019. Low-light Image Enhancement of High-speed Endoscopic Videos using a Convolutional Neural Network. Medical Biol. Eng. Comput. 57(7): 1451-1463.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al., 2014. Generative Adversarial Networks. In: International Conference on Neural Information Processing Systems (NIPS). pp. 2672–2680.

Grubinger, M., Clough, P., Müller, H., Deselaers, T., 2006. The IAPR TC12 Benchmark: A New Evaluation Resource for Visual Information Systems. Int. Workshop OntoImage, 5.

Gu, K., Zhai, G., Lin, W., Liu, M., 2016. The Analysis of Image Contrast: From Quality Assessment to Automatic Enhancement. IEEE Trans. Cybern. 46, 284–297.

Guan, J., Lai, R., Xiong, A., 2019. Wavelet Deep Neural Network for Stripe Noise Removal. IEEE Access, 7, 44544-44554.

Guo, C., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., Cong, R., 2020. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In: IEEE Confernence on Computer Vision and Pattern Recognition (CVPR). pp. 1777-1786.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M. S., 2016. Deep Learning for Visual Understanding: A Review. Neurocomouting 187, 27- 48.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: IEEE International Conference on Computer Vision (ICCV). pp. 2980-298.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770-778.

Hojatollah, Y., Wang, Z., 2013. Objective Quality Assessment of Tone-mapped Images. IEEE Trans. Image Process. 22, 657-667.

Hua, W., Xia, Y., 2018. Low-Light Image Enhancement Based on Joint Generative Adversarial Network and Image Quality Assessment. In: International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). pp. 1-6.

Huang, J.-B., Singh, A., Ahuja, N., 2015. Single Image Super-Resolution from Transformed Self-Exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5197-5206.

Huang, S.-C., Cheng, F.-C., Chiu, Y.-S., 2013. Efficient Contrast Enhancement Using Adaptive Gamma Correction with Weighting Distribution. IEEE Trans. Image Process. 22, 1032-1041.

Jiang, L., Jing, Y., Hu, S., Ge, B., Xiao, W., 2018. Deep Refinement Network for Natural Low-Light Image Enhancement in Symmetric Pathways. Symmetry 10, 491.

Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., et al., 2021. EnlightenGAN: Deep Light Enhancement without Paired Supervision. IEEE Trans. Image Process. 30, 2340-2349.

Jobson, D. J., Rahman, Z., Woodel, G. A., 1997a. Properties and Performance of a Center/Surround Retinex. IEEE Trans. Image Process. 6, 451-462.

Jobson, D. J., Rahman, Z., Woodell, G. A., 1997b. A Multsiscale Retinex for Bridging the Gap Between Color Images and the Human Observation of Scenes. IEEE Trans. Image Process. 6, 965-976.

Jolicoeur-Martineau, A., 2018. The Relativistic Discriminator: A Key Element Missing from Standard GAN. ArXiv preprint arXiv:1807.00734.

Kalantari, N. K., Ramamoorthi, R., 2017. Deep High Dynamic Range Imaging of Dynamic Scenes. ACM Trans. Graph. 36, 144:1–144:12.

Khan, R., Yang, Y., Liu, Q., Qaisar, Z.H., 2021a. A Ghostfree Contrast Enhancement Method for Multiview Images Without Depth Information. J. Vis. Commun. Image Represent.,78.

Khan R., Yang Y., Liu Q., Shen J., and Li B., , 2021d. Deep Image Enhancement for Ill-Light Imaging. Journal of the Optical Society of America A, 38(6): 827-839

Khan, R., Liu, Q., Yang, Y., 2021c. A Deep Hybrid Few Shot Divide and Glow Method for Ill-Light Image Enhancement. IEEE Access, 9, 17767-17778.

Khan, R., Yang, Y., Liu, Q., Qaisar, Z.H., 2021d. Divide and conquer: Ill-light Image Enhancement via Hybrid Deep Network. Expert Systems with Applications, 182.

Kim, G., Kwon, D., Kwon, J., 2019. Low-Lightgan: Low-Light Enhancement via Advanced Generative Adversarial Network with Task-Driven Training. In: IEEE International Conference on Image Processing (ICIP). pp. 2811-2815.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: Neural Information Processing Systems Conference (NIPS). pp. 1106-1114.

Land, E., McCann, J., 1971. Lightness and Retinex Theory. J. Opt. Soc. Amer. 61, 1-11.

Han, L., Xiong, J., Geng, G., Zhou, M., 2009. Using HSV Space Real-color Image Enhanced by Homomorphic Filter in Two Channels. Comput. Eng. Appl., 45(27), 18-20.

Hao, S., Han, X., Guo, Y., Xu, X., Wang, M., 2020. Low-Light Image Enhancement with Semi-Decoupled Decomposition. IEEE Transactions on Multimedia, 22, 3025-3038.

Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Ng, A. Y., 2011. On Optimization Methods for Deep Learning. In: International Conference on Machine Learning (ICML). pp. 265–272.

Lee, H.-G., Yang, S., Sim, J.-Y., 2015. Color Preserving Contrast Enhancement for Low-light Level Images based on Retinex. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). pp. 884-887.

Li, C., Guo, J., Porikli, F., Pang, Y., 2018. LightenNet: A Convolutional Neural Network for Weakly Illuminated Image Enhancement. Pattern Recognit. Lett. 104, 15-22.

Li, C., Anwar, S., Porikli, F., 2020. Underwater Scene Prior Inspired Deep Underwater Image and Video Enhancement. Pattern Recognit., 98.

Li, L., Wang, R., Wang, W., Gao, W., 2015. A Low-light Image Enhancement Method for Both Denoising and Contrast Enlarging. In: IEEE International Conference on Image Processing (ICIP). pp. 3730-3734.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal Loss for Dense Object Detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 2999-3007.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., et al., 2014. Microsoft COCO: Common Objects in Context. In: European Conference on Computer Vision (ECCV). pp. 740-755.

Liu, X., Tanaka, M., Okutomi, M., 2012. Noise Level Estimation Using Weak Textured Patches of a Single Noisy Image. In: IEEE International Conference on Image Processing (ICIP). pp. 665-668.

Loh, Y. P., Chan, C. S., 2019. Getting to Know Low-light Images with the Exclusively Dark Dataset. Comput. Vision Image Understanding 178, 30-42.

Lore, K. G., Akintayo, A., Sarkar, S., 2017. LLNet: A Deep Autoencoder Approach to Natural Low-light Image Enhancement. Pattern Recognit. 61, 650-662.

Lv, F., Li, Y., Lu, F., 2020. Attention Guided Low-light Image Enhancement with a Large Scale Low-light Simulation Dataset. ArXiv preprint arXiv:1908.00682.

Lv, F., Lu, F., Wu, J., Lim, C., 2018. MBLLEN: Low-light Image/Video Enhancement Using CNNs. In: British Machine Vision Conference (BMVC). pp. 220.

McGill, M., 1983. Introduction to Modern Information Retrieval. NewYork: McGrawHill.

Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., Zhang, L., 2017. Waterloo Exploration Database: New Challenges for Image Quality Assessment Models. IEEE Trans. Image Process. 26, 1004-1016.

Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z., 2022. Toward Fast, Flexible, and Robust Low-Light Image Enhancement. ArXiv preprint arXiv:2204.10137.

Ma Y., Liu J., Liu Y., Fu H., Hu Y., Cheng J., Qi H., Wu Y., Zhang J., and Zhao Y., 2021. Structure and Illumination Constrained GAN for Medical Image Enhancement. IEEE Transactions on Medical Imaging 40(12): 3955-3967.

Meng, Y., Kong, D., Zhu, Z., Zhao, Y., 2019. From Night to Day: GANs Based Low Quality Image Enhancement. Neural Process. Lett. 50, 799-814.

Milford, M. J., Wyeth, G. F., 2012. SeqSLAM: Visual Route-based Navigation for Sunny Summer Days and Stormy Winter Nights. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 1643-1649.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2020. Image Segmentation Using Deep Learning: A Survey. ArXiv preprint arXiv:2001.05566.

Mittal, A., Moorthy, A. K., Bovik, A. C., 2012. No-Reference Image Quality Assessment in the Spatial Domain. IEEE Trans. Image Process. 21, 4695-4708.

Mittal, A., Soundararajan, R., Alan C. Bovik., 2013. Making a 'Completely Blind' Image Quality Analyzer. IEEE Signal Process. Lett. 20, 209-212.

Murray, N., Marchesotti, L., Perronnin, F., 2012. AVA: A Large-scale Database for Aesthetic Visual Analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2408-2415.

Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and Checkerboard Artifacts. Distill, 1.

Pisano, E. D., Zong, S., Hemminger, B. M., DeLuca, M., Johnston, R. E., Muller, K., et al., 1998. Contrast Limited Adaptive Histogram Equalization Image Processing to Improve the Detection of Simulated Spiculations in Dense Mammograms. J. Digit. Imaging 11, 193-200.

Rahman, Z., Jobson, D. J., Woodell, G. A., 1996. Multi-scale Retinex for Color Image Enhancement. In: IEEE International Conference on Image Processing (ICIP). pp. 1003-1006.

Ranzato, M., Poultney, C., Chopra, S., LeCun, Y., 2006. Efficient Learning of Sparse Representations with an Energy-based Model. In Neural Information Processing Systems (NIPS). pp. 1137-1144.

Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement. ArXiv preprint arXiv:1804.02767.

Ren, W., Liu, S., Ma, L., Xu, Q., Xu, X., 2019. Low-Light Image Enhancement via a Deep Hybrid Network. IEEE Trans. Image Process. 28, 4364-4375.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiv preprint arXiv:1505.04597.

Salem, C., Azar, D., Tokajian, S., 2018. An Image Processing and Genetic Algorithm-Based Approach for the Detection of Melanoma in Patients. Methods Inf. Med. 57, 74-80 .

Schaefer, G., Stich, M., 2003. UCID: An Uncompressed Color Image Database. In: Storage and Retrieval Methods and Applications for Multimedia. pp. 472-480.

Schölkopf, B., Smola, A. J., Williamson, R. C., Bartlett, P. L., 2000. New Support Vector Algorithms. Neural Comput. 12, 1207-1245.

Schwartz, E., Giryes, R., Bronstein, A. M., 2019. DeepISP: Toward Learning an End-to-End Image Processing Pipeline. IEEE Trans. Image Process. 28, 912-923.

Sheikh, H. R., Bovik, A. C., 2006. Image Information and Visual Quality. IEEE Trans. Image Process. 15, 430-444.

Shen, L., Yue, Z., Feng, F., Chen, Q., Liu, S., Ma, J., 2017. MSR-net: Low-light Image Enhancement Using Deep Convolutional Network. ArXiv preprint arXiv:1711.02488.

Sheng, J., Lv, G., Du, G., Wang, Z., Feng, Q., 2022. Multi-scale Residual Attention Network for Single Image Dehazing. Digital Signal Processing, 121.

Shin, Y.-G., Sagong, M.-C., Yeo, Y.-J., Ko, S.-J., 2018. Adversarial Context Aggregation Network for Low-Light Image Enhancement. In: Digital Image Computing: Techniques and Applications (DICTA). pp. 1-5.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv preprint arXiv:1409.1556.

Su, Y., Lian, Q., Zhang, X., Shi, B., Fan, X., 2019. Multi-scale Cross-path Concatenation Residual Network for Poisson Denoising. IET Image Processing, 13(8), 1295-1303.

Sun, L., Hays, J., 2012. Super-resolution from Internet-scale Scene Matching. In: IEEE International Conference on Computational Photography (ICCP). pp. 1-12.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818-2826.

Talebi, H., Milanfar, P., 2018. Nima: Neural Image Assessment. IEEE Trans. Image Process. 27, 3998–4011.

Tao, L., Zhu, C., Xiang, G., Li, Y., Jia, H., Xie, X., 2017. LLCNN: A Convolutional Neural Network for Low-light Image Enhancement. In: IEEE Visual Communications and Image Processing (VCIP). pp. 1-4.

Tian, C., Xu, Y., Zuo, W., 2020. Image Denoising Using Deep CNN with Batch Renormalization. Neural Networks, 121, 461-473.

Ting, F., Fuquan, Z., Zhaochai, Y., Zuoyong, L., 2022. Image Dehazing Network Based on Multi-scale Feature Extraction. In: Proceedings of Advances in Smart Vehicular Technology, Transportation, Communication and Applications. pp. 391-399.

VidalMata, R. G., Banerjee, S., RichardWebster, B., Albright, M., Davalos, P., McCloskey, S., et al., 2020. Bridging the Gap Between Computational Photography and Visual Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 1-1.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and Composing Robust Features with Denoising Autoencoders. In: International Conference on Machine Learning (ICML). pp. 1096–1103.

Wang, J., Tan, W., Niu, X., Yan, B., 2019. RDGAN: Retinex Decomposition Based Adversarial Learning for Low-Light Enhancement. In: IEEE International Conference on Multimedia and Expo (ICME). pp. 1186-1191.

Wang, L., Fu, G., Jiang, Z., Ju, G., Men, A., 2019. Low-Light Image Enhancement with Attention and Multi-level Feature Fusion. In: IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 276-281.

Wang, R., Zhang, Q., Fu, C.-W., Shen, X., Zheng, W.-S., Jia, J., 2019. Underexposed Photo Enhancement Using Deep Illumination Estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6842-6850.

Wang, S., Zheng, J., Hu, H.-M., Li, B., 2013. Naturalness Preserved Enhancement Algorithm for Non-Uniform Illumination Images. IEEE Trans. Image Process. 22, 3538-3548.

Wang, W., Wei, C., Yang, W., Liu, J., 2018. GLADNet: Low-Light Enhancement Network with Global Awareness. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG). pp. 751-755.

Wang, Y., Chen, Q., Zhang, B., 1999. Image Enhancement Based on Equal Area Dualistic Sub-image Histogram Equalization Method. IEEE Trans. Consum. Electron. 45, 68-75.

Wang, Z., Li, Q., 2011. Information Content Weighting for Perceptual Image Quality Assessment. IEEE Trans. Image Process. 20, 1185-1198.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E., 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Trans. Image Process. 13, 600–612.

Wang, Z., Simoncelli, E. P., Bovik, A.C., 2003. Multi-Scale Structural Similarity for Image Quality Assessment. In: IEEE Asilomar Conference on Signals, Systems, and Computers. pp. 1398-1402.

Wei Liu, D. A., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C., 2016. SSD: Single Shot MultiBox Detector. In: European Conference on Computer Vision (ECCV). pp. 21-37.

Wei, C., Wang, W., Yang, W., Liu, J., 2018. Deep Retinex Decomposition for Low-Light Enhancement. In: British Machine Vision Conference (BMVC). pp. 155.

Xiang, Y., Fu, Y., Zhang, L., Huang, H., 2019. An Effective Network with ConvLSTM for Low-Light Image Enhancement. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 221-233.

Xiao, J., Song, S., Ding, L., 2008. Research on the fast algorithm of spatial homomorphic filtering, (in Chinese). J. Image Graph., 13 (12), 2302–2306.

Xu, L., Wei, Y., 2022. Pyramid Deep Dehazing: An Unsupervised Single Image Dehazing Method Using Deep Image Prior. Optics & Laser Technology, 148.

Xu, W., Lee, M., Zhang, Y., You, J., Suk, S., Choi, J.-y., 2018. Deep Residual Convolutional Network for Natural Image Denoising and Brightness Enhancement. In: International Conference on Platform Technology and Service (PlatCon). pp. 1-6.

Yang, W., Yuan, Y., Ren, W., Liu, J., Scheirer, W. J., Wang, Z., Zhang, T., 2020. Advancing Image Understanding in Poor Visibility Environments: A Collective Benchmark Stud. IEEE Trans. Image Process. 29, 5737-5752.

Yangming, S., Xiaopo, W., Ming, Z., 2019. Low-light Image Enhancement Algorithm Based on Retinex and Generative Adversarial Network. ArXiv preprint arXiv:1906.06027.

Zaheeruddin, S., Suganthi, K., 2019. Image Contrast Enhancement By Homomorphic Filtering based Parametric Fuzzy Transform. In: Proc. Comput. Sci. vol. 165, pp. 166–172.

Zeiler, M. D., Fergus, R., 2014. Visualizing and Understanding Convolutional Networks. In: European Conference on Computer Vision (ECCV). pp. 818-833.

Zhang, C., Liu, W., Xing, W., 2018. Color Image Enhancement based on Local Spatial Homomorphic Filtering and Gradient Domain Variance Guided Image Filtering. J. Electron. Imaging, 27(6).

Zhang, L., Zhang, L., Mou, X., Zhang, D., 2011. Fsim: A Feature Similarity Index for Image Quality Assessment. IEEE Trans. Image Process. 20, 2378-2386.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586-595.

Zhang, Y., Di, X., Zhang, B., Wang, C., 2020. Self-supervised Image Enhancement Network: Training with Low-light Images Only. ArXiv preprint arXiv:2002.11300.

Zhang, Y., Zhang, I., Guo, X., 2019. Kindling the Darkness: A Practical Low-light Image Enhancer. In: ACM International Conference on Multimedia (ACM MM). pp. 1632–1640.

Zhang, Y., Xie, M., 2013. Colour Image Enhancement Algorithm based on HSI and Local Homomorphic Filtering (in Chinese). Comput. Appl. Softw., 30(12), 303–307.

Zhi, N., Mao, S., Li, M., 2018. An Enhancement Algorithm for Coal Mine Low Illumination Images Based on Bi-Gamma Function. J. Liaoning Tech. Univ. 37, 191-197.

Zosso, D., Tran, G., Osher, S., 2015. Non-Local Retinex - A Unifying Framework and Beyond. SIAM J. Imaging Sci., 8, 787-826.

Appendix A

We show below the boxplots for the distribution of the IQA metrics computed for the ExDark dataset, and ranked from best to worst following the average measure. In addition, we include the standard deviation and number of outliers for each of the metrics.

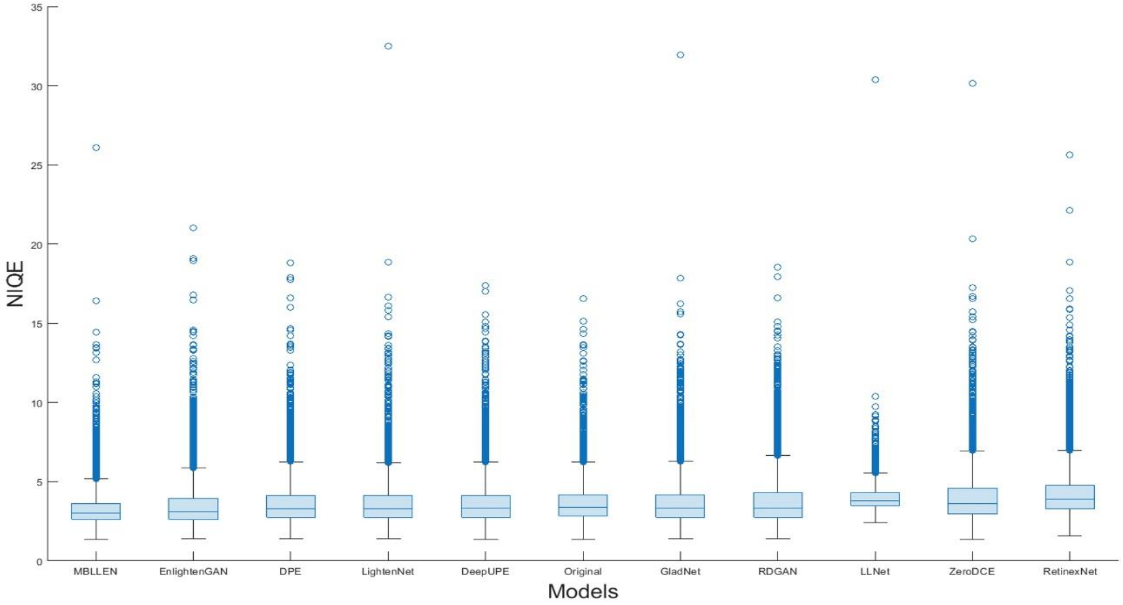


Fig. A.1. Boxplots for NIQE measures for the ExDark dataset

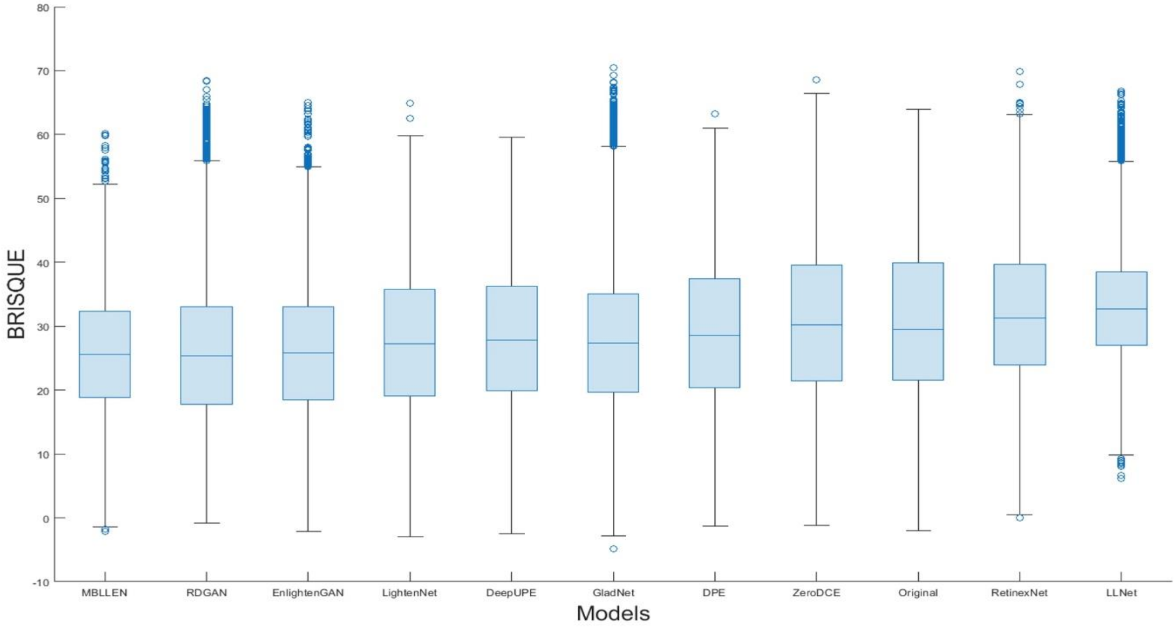


Fig. A.2. Boxplots for BRISQUE measures for the ExDark dataset

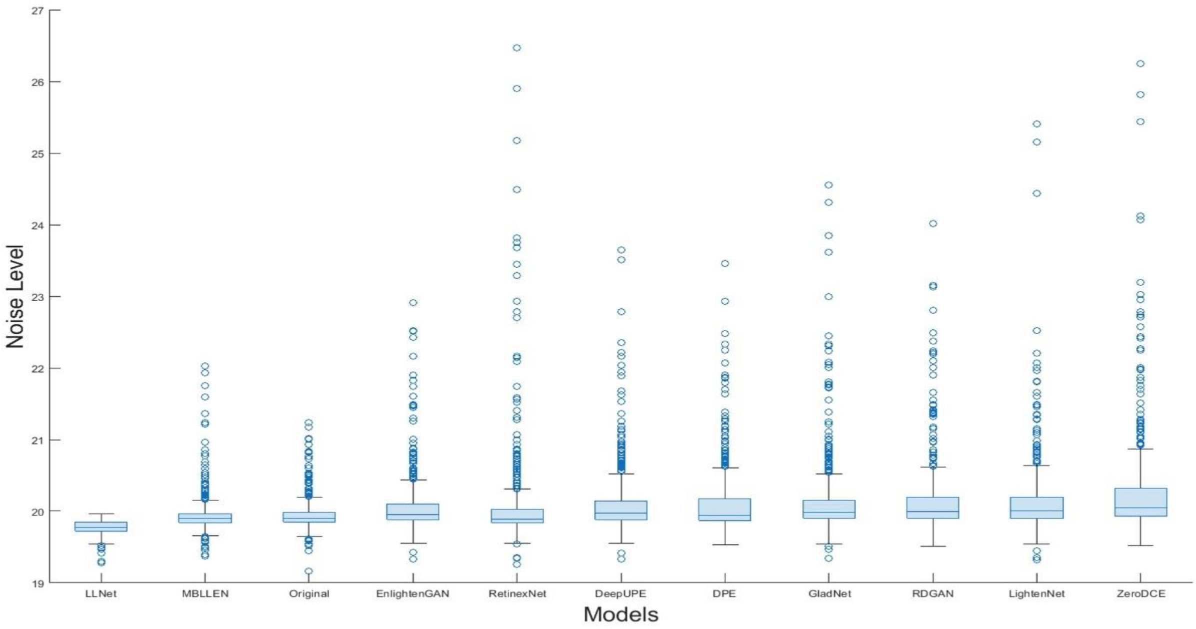


Fig. A.3. Boxplots for Noise Level measures for the ExDark subset

Table A.1. Standard deviation and number of outliers for the distributions of the IQA metric measures reported for the original and enhanced images of the ExDark dataset.

a. Standard deviation results

Approach	*NIQE*	*BRISQUE*	*Noise*
Original LLIs	1.35	12.13	0.21
MBLLEN	1.16	10.10	0.27
EnlightenGAN	1.49	11.11	0.41
DPE	1.44	12.03	0.41
LightenNet	1.51	11.35	0.56
DeepUPE	1.46	11.14	0.46
GladNet	1.63	12.17	0.56
RDGAN	1.81	12.10	0.50
LLNet	0.84	9.75	0.09
ZeroDCE	1.71	11.96	0.75
RetinexNet	1.64	11.38	0.74

b. Number of outliers

Approach	*NIQE*	*BRISQUE*	*Noise*
Original LLIs	332	0	47
MBLLEN	390	24	61
EnlightenGAN	486	62	55
DPE	367	1	51
LightenNet	421	2	49
DeepUPE	374	0	56
GladNet	614	234	57
RDGAN	625	201	43
LLNet	412	252	7
ZeroDCE	493	1	53
RetinexNet	448	13	67

Appendix B

We show below the boxplots for the distribution of the IQA metrics computed for the LOL dataset, and ranked from best to worst following the average measure. In addition, we include the standard deviation and number of outliers for each of the metrics.

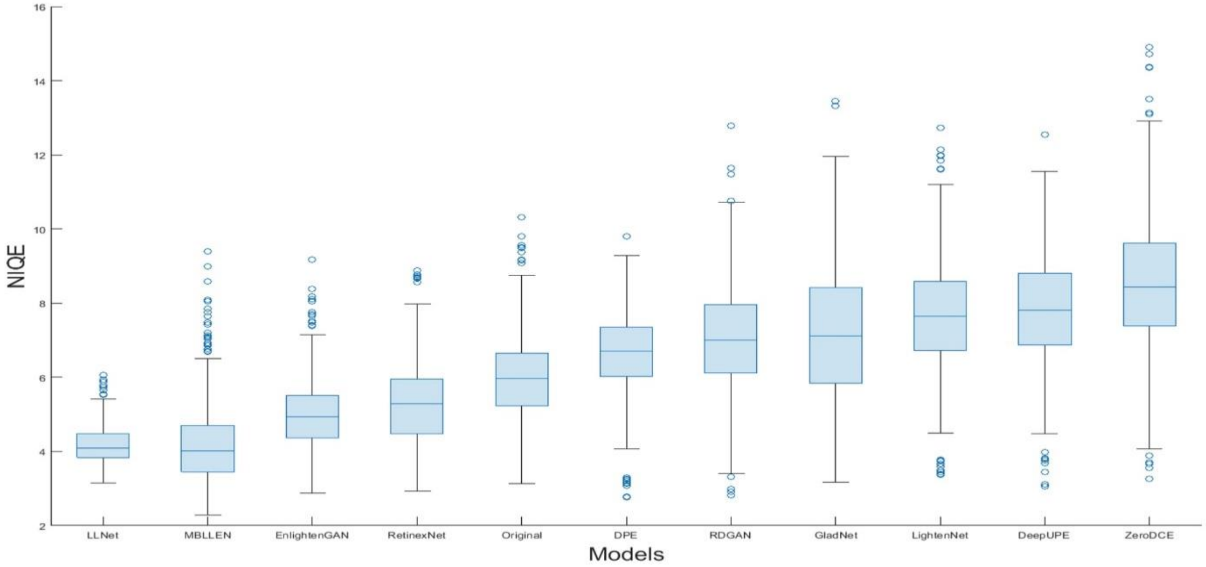


Fig. B.1. Boxplots for NIQE measures for the LOL dataset

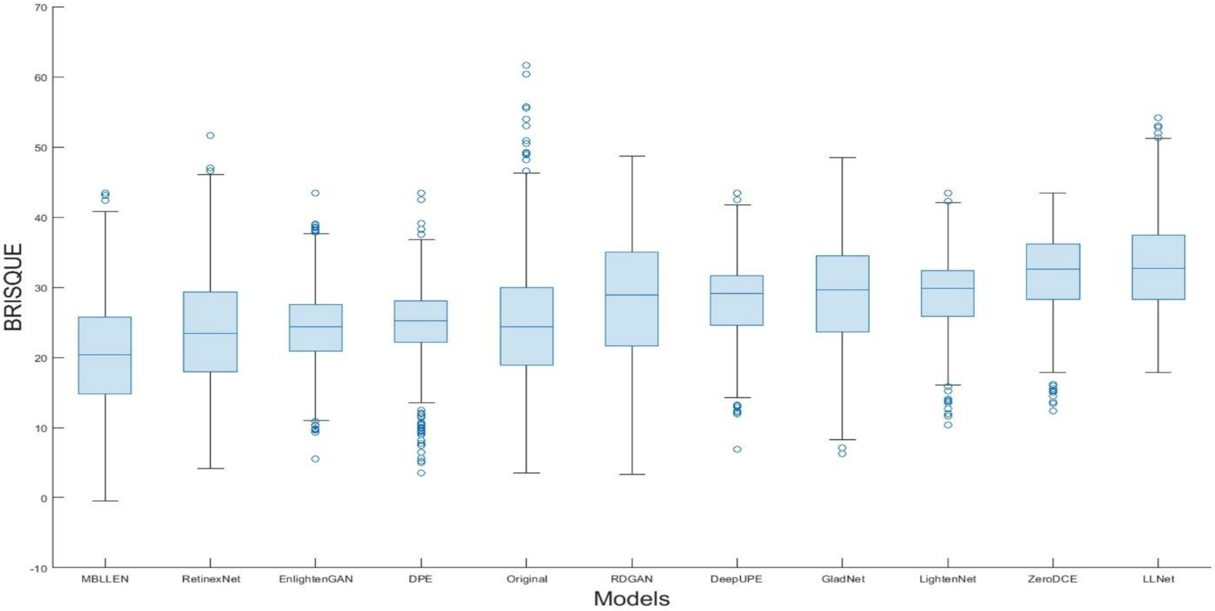


Fig. B.2. Boxplots for BRISQUE measures for the LOL dataset

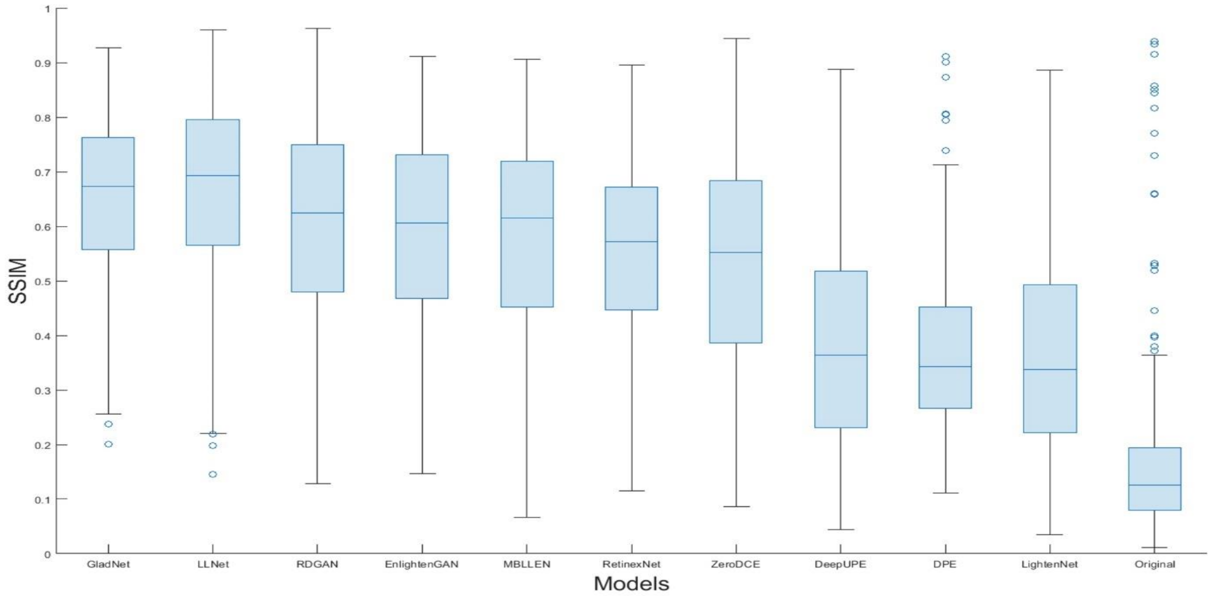


Fig. B.3. Boxplots for SSIM measures for the LOL dataset

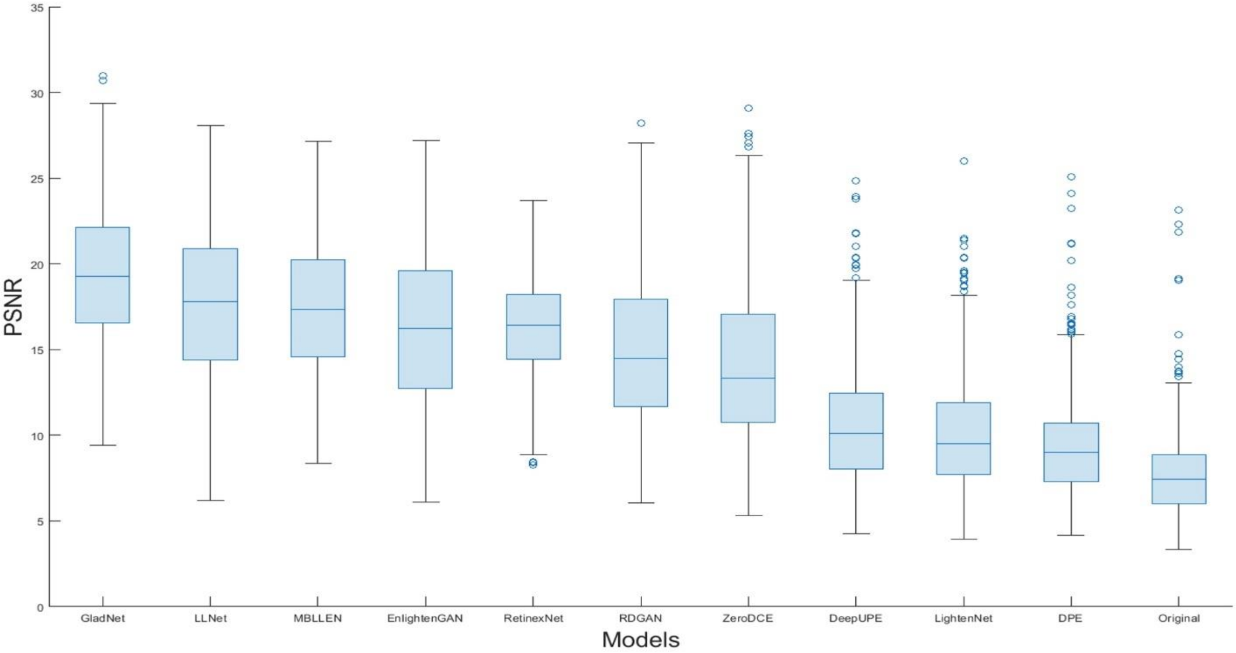


Fig. B.4. Boxplots for PSNR measures for the LOL dataset

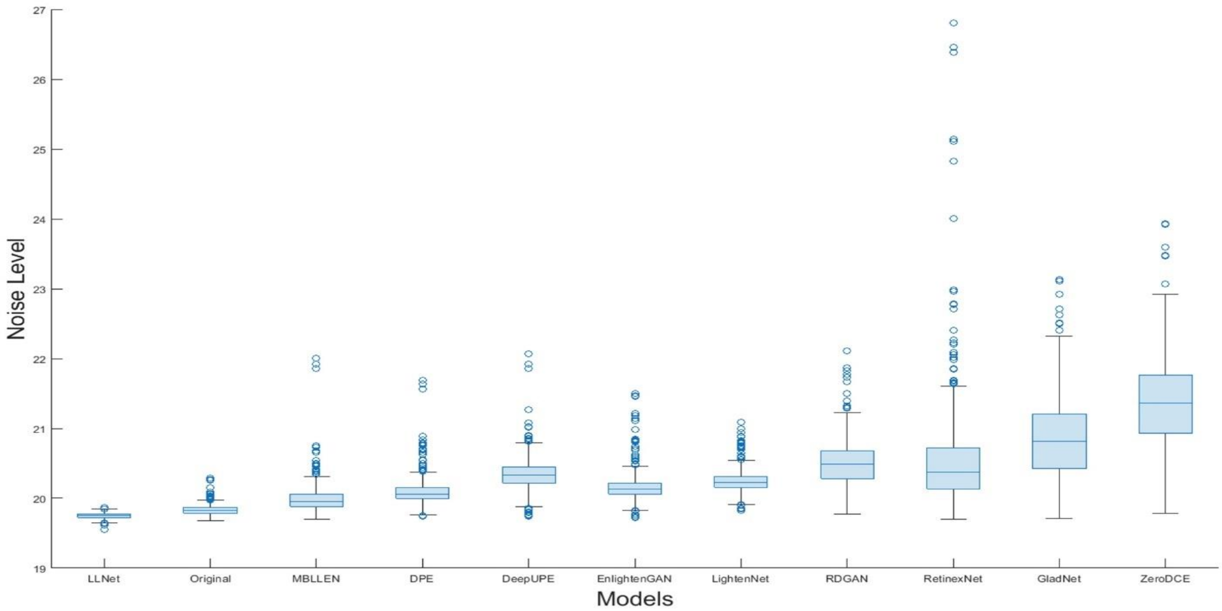


Fig. B.5. Boxplots for Noise Level measures for LOL dataset

Table B.1. Standard deviation and number of outliers for the distributions of the IQA metric measures reported for the original and enhanced images of the LOL dataset.

a. Standard deviation results

Approach	*NIQE*	*BRISQUE*	*SSIM*	*PSNR*	*Noise*
Original	1.15	9.22	0.13	2.58	0.07
MBLLEN	1.12	7.82	0.176	3.92	0.217
EnlightenGAN	0.93	5.88	0.174	4.64	0.213
DPE	1.12	5.74	0.14	3.07	0.20
LightenNet	1.499	5.40	0.18	3.41	0.16
DeepUPE	1.490	5.62	0.19	3.51	0.24
GladNet	1.77	7.84	0.15	3.86	0.58
RDGAN	1.50	8.71	0.177	4.48	0.33
LLNet	0.50	6.33	0.170	4.24	0.03
ZeroDCE	1.80	5.72	0.18	4.70	0.65
RetinexNet	1.11	9.08	0.15	2.90	0.81

b. Number of outliers results

Approach	*NIQE*	*BRISQUE*	*SSIM*	*PSNR*	*Noise*
Original	9	14	21	13	21
MBLLEN	23	4	0	0	18
EnlightenGAN	12	19	0	0	33
DPE	10	32	8	21	31
LightenNet	16	16	0	14	37
DeepUPE	10	13	0	12	24
GladNet	2	2	2	2	8
RDGAN	8	0	0	1	13
LLNet	9	5	3	0	6
ZeroDCE	15	10	0	5	6
RetinexNet	7	3	0	3	30

Appendix C

Figures C.1 to C.4. show the precision-recall (*P-R*) curves for the 12 classes of the ExDark dataset obtained by the four detection models considered in our experimental study, applied on the original LLIs and their enhanced counterparts. Average precision (*AP*) results per class (shown in the legend) are ranked from best to worst following each of the enhancement models.

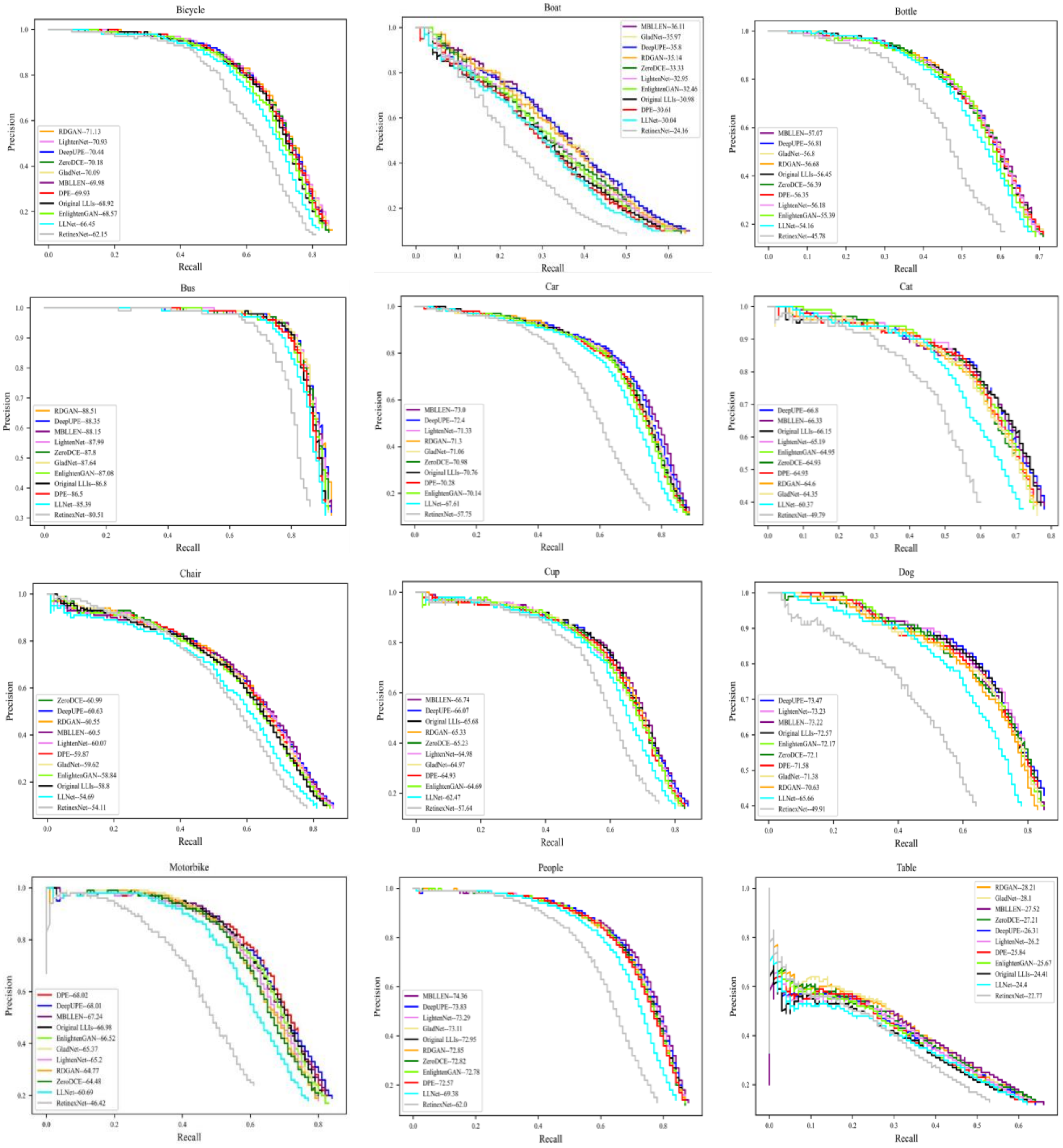


Fig. C.1. P-R Curves for YOLOv3

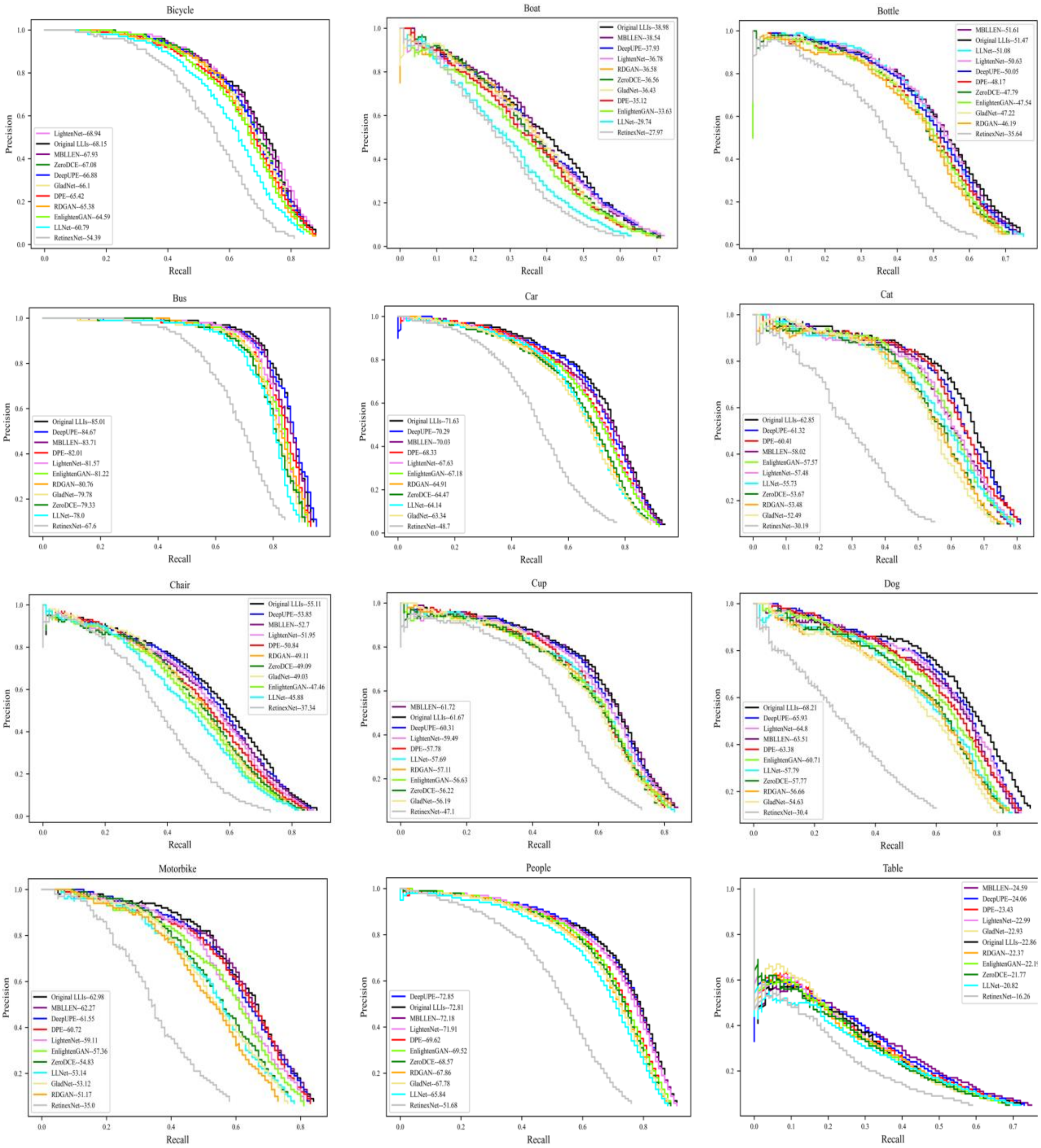


Fig. C.2. P-R Curves for RetinaNet

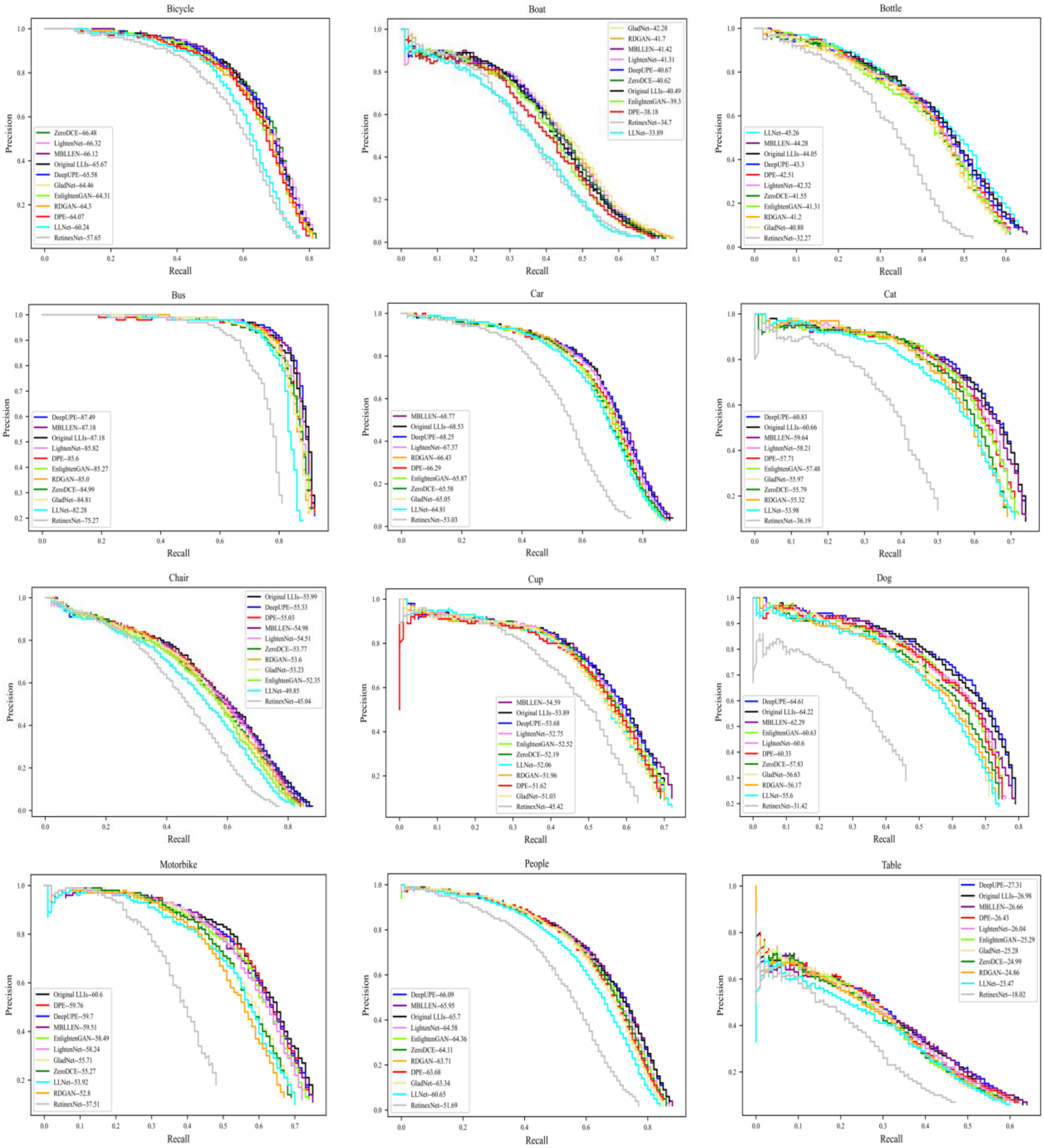


Fig. C.3. P-R Curves for SSD

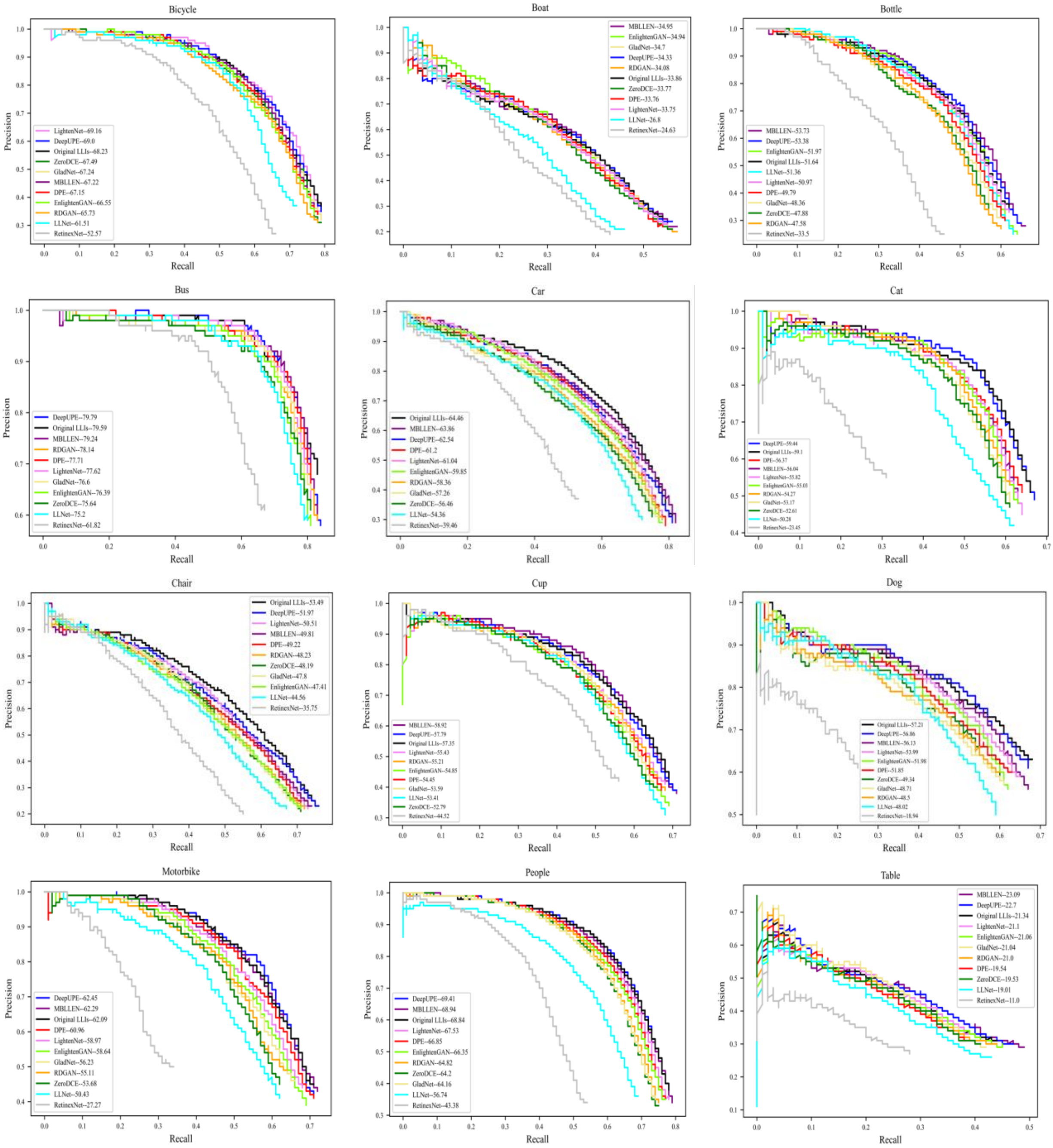


Fig. C.4. P-R Curves for MaskRCNN