Lightweight HDR Camera ISP for Robust Perception in Dynamic Illumination Conditions via Fourier Adversarial Networks

Pranjay Shyam[1]
pranjayshyam@kaist.ac.kr
Sandeep Singh Sengar[2]
sengar@di.ku.dk
Kuk-Jin Yoon[1]
kjyoon@kaist.ac.kr
Kyung-Soo Kim[1]
kyungsookim@kaist.ac.kr

[1] Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea

[2] University of Copenhagen
Copenhagen, Denmark

Abstract

The limited dynamic range of commercial compact camera sensors results in an inaccurate representation of scenes with varying illumination conditions, adversely affecting image quality and subsequently limiting the performance of underlying image processing algorithms. Current state-of-the-art (SoTA) convolutional neural networks (CNN) are developed as post-processing techniques to independently recover under-/over-exposed images. However, when applied to images containing real-world degradations such as glare, high-beam, color bleeding with varying noise intensity, these algorithms amplify the degradations, further degrading image quality. We propose a lightweight two-stage image enhancement algorithm sequentially balancing illumination and noise removal using frequency priors for structural guidance to overcome these limitations. Furthermore, to ensure realistic image quality, we leverage the relationship between frequency and spatial domain properties of an image and propose a Fourier spectrum-based adversarial framework (AFNet) for consistent image enhancement under varying illumination conditions. While current formulations of image enhancement are envisioned as post-processing techniques, we examine if such an algorithm could be extended to integrate the functionality of the Image Signal Processing (ISP) pipeline within the camera sensor benefiting from RAW sensor data and lightweight CNN architecture. Based on quantitative and qualitative evaluations, we also examine the practicality and effects of image enhancement techniques on the performance of common perception tasks such as object detection and semantic segmentation in varying illumination conditions.

1 Introduction

Images captured in dynamic illumination conditions can have underexposed or overexposed regions or a combination of both. The underexposed regions are susceptible to noise, and overexposed regions obscure textural information of surrounding features. This deteriorates the performance of underlying high-level perception tasks such as feature matching [33],

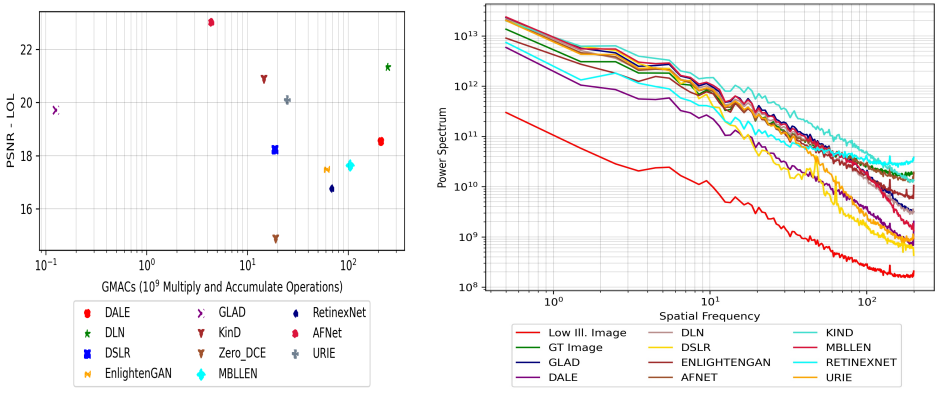

Figure 1: Performance landscape (GMACs vs PSNR) of different SoTA Image Enhancement Algorithms on sRGB Images from the LOL [44] dataset (left) and corresponding Power Spectral Density curves of enhanced images (right).

lane detection [22], object detection [27], and semantic segmentation [38]. While hardware modifications or software adjustments can be used for increasing light received by a camera sensor when capturing a scene, these approaches introduce additional noise and artifacts such as motion blur (increasing exposure time), losing the depth of field (increasing aperture of the appropriate lens), and non-uniform lightening (using additional light source). Hence focus shifts towards software-based image enhancement as a post-processing technique to enhance image quality while maintaining image sharpness and color balance.

Current SoTA algorithms leverage CNNs and define different functional configurations focusing on CNN architectures [24, 48, 51] or optimization formulation [12, 52] to obtain a well-illuminated image in low light or high illumination conditions. However, illumination settings confine the performance of these methods; hence they perform well only in the conditions wherein the complete image has similar illumination conditions. This assumption is rarely fulfilled in real scenarios, resulting in increased pixel noise, color bleeding, and pixelations, reducing image quality when using these algorithms on natural images containing local illumination sources. This is extremely detrimental in scenarios wherein these enhanced images are used as inputs for performing high-level vision tasks such as object detection, semantic segmentation, etc., as it degrades the performance of SoTA algorithms (See the supplementary).

To circumvent these limitations of SoTA low-light image enhancement (LLIE) algorithms, we propose a two-stage enhancement architecture wherein the first stage focuses on coarsely balancing illumination and the second stage focuses on noise and artifact removal to reconstruct a well-illuminated color-balanced image. Furthermore, to improve feature quality without increasing computations, we propose a compact multi-scale feature extraction mechanism that splits a given feature map along channel dimension and subsequently uses a convolutional filter with different kernel sizes. These features are then aggregated after being scaled using a channel attention mechanism that encourages relevant features while suppressing irrelevant features, allowing us to obtain features across diverse receptive fields used to balance the illumination of the image. Subsequently, the secondary network is used to recover the regions affected by noise and artifacts to ensure textural and structural fidelity within enhanced images. This two-stage approach reduces the network size while ensuring SoTA performance, saving on inference time and memory requirement.

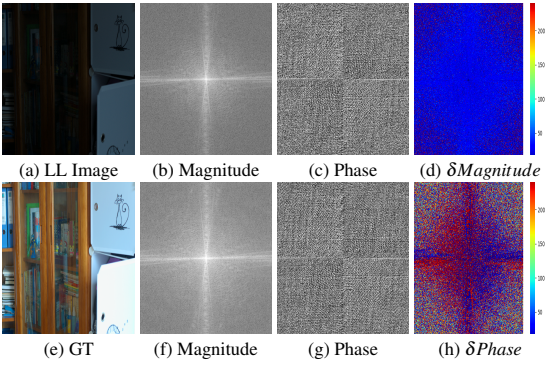

Figure 2: Disentangling (a) low light and (e) its corresponding ground truth into (b, f) magnitude and (c, g) phase components using fast Fourier transform with difference heatmap of magnitude and phase (d, h). In the heatmap, red highlights maximum error whereas blue represents minimum error.

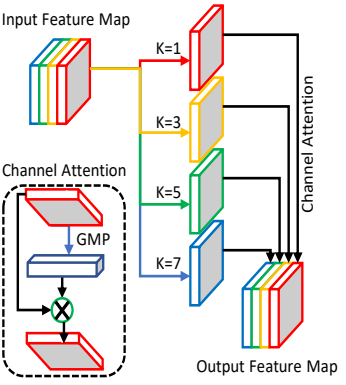

Figure 3: Overview of the proposed compact multiscale-feature extraction and aggregation.

Furthermore, upon a closer inspection of power spectral density of images enhanced by SoTA algorithms, we observe poor performance at high frequencies that capture edge information; thus, we propose utilizing the frequency domain information to ensure consistent enhancement under diverse conditions by leveraging the duality between frequency and spatial domain characteristics of an image. Specifically, point-wise modifications in the frequency domain result in global modifications across spatial domains in an image. In addition, visual examination (Fig. 2) of frequency domain information, i.e., magnitude and phase components generated using Fast Fourier Transform (FFT), reveals multiple attributes that can be leveraged to ensure image enhancement. Some notable attributes include the presence of high textural information within a well-lit image (Fig. 2(e)) that are centered in the magnitude spectrum. As low light image doesn't capture detailed textural information, the intensity of magnitude spectrum (Fig. 2(b)) is attenuated. This observation can be extended for edges present in an image. While the magnitude spectrum can be interpreted as 'how much' of frequencies are present in an image, an equally important phase component (Fig. 2(c, f)) determines 'where' those frequencies are present in the image. This motivates us to construct an adversarial network that utilizes both the magnitude and phase components of an image to determine whether it is real/fake. Such a binary CNN would leverage the complete spectral properties of how much and where certain frequencies are present and thus result in enhanced images closely resembling the ground truth.

While the performance of image enhancement algorithms has improved lately, utilization of camera-ISP (comprising of multiple handcrafted task-specific algorithms to convert raw color filter array (CFA) data to standard RGB (sRGB) image) introduces additional non-linearities capping the peak performance of SoTA algorithms. However, due to the proprietary nature of camera-ISP, the implementation of enhancement algorithms on RAW sensor data is not well studied. Recently different works such as SID [4], and Five5K [2] have collected RAW image pairs using different exposure times to construct paired images that could be used for training end-to-end image enhancement algorithms instead of the current post-processing formulation. However, as these datasets capture high illumination conditions by increasing the exposure times, they do not contain dynamic scenarios such as glares, color bleeding, high beam, etc., thus representing conditions that are easily violated in real-life deployment. Nevertheless, they could be used for training an end-to-end CNN with additional

augmentations as paired-dataset capturing diverse illumination conditions and storing it in RAW format is impractical. Instead, we construct a test-set that represents varying illumination conditions under diverse conditions by capturing images using a dashboard camera mounted on a consumer vehicle, allowing us to examine the efficacy of algorithms on real-world deployment. We summarize our contributions as,

- We propose a two-stage CNN architecture for performing illumination balancing and image restoration that works with both sRGB and RAW images.
- We combine channel split mechanism with multiscale convolutions to enhance the receptive field and increase feature information without a substantial increase in computations.
- To ensure the presence of high-frequency components within enhanced images, we propose using frequency information within an adversarial learning mechanism.
- As paired training datasets cannot represent dynamic conditions, we construct an unpaired test-set by collecting RAW and sRGB images under dynamic illumination conditions using a personal vehicle as data capturing setup.
- We demonstrate varying illumination conditions to adversely affect the performance of object detection algorithms and improve it by enhancing image quality using the proposed approach.

2 Related Works

2.1 Image Enhancement

Early CNN-based approach, LLNet [24], proposed an autoencoder formulation for performing contrast enhancement while simultaneously suppressing noise. Subsequent works rely on Retinex Theory coupled with additional priors such as structure aware loss in RetinexNet [44], reflectance restoration in KinD [51], and attention mechanism [48] to improve LLIE performance. When applying these algorithms on images comprising both well and poorly-lit regions, they distort the regions that do not require any enhancement. To overcome such situations, DALE [15] proposed a two-stage approach of first identifying dark regions using a visual attention module and then enhancing the brightness of these regions. These methods perform LLIE on images of reduced spatial resolution resulting in inaccurate spatial enhancement. MIRNet [47] was proposed to maintain a semantically and spatially accurate enhancement network using multi-resolution convolution and attention mechanisms. As these methods require paired training samples, constructing a training dataset is extremely time-consuming. EnlightenGAN [12] utilized a generative adversarial framework for constructing low light images and subsequently using it to train an underlying LLIE algorithm, whereas [52] relying on self-supervised learning to formulate a retinex model optimized using maximum entropy.

Recently frequency priors have been explored to restore images with MWCNN [21] using wavelet transforms to perform tasks such as super-resolution, denoising, and JPEG artifact removal. [46] highlighted that detecting and removing noise is much easier from low-frequency components and thus proposed a two-stage network that decomposes an image into low and high frequency, recovers low-frequency components, and enhances high-frequency details. In addition, DIDH [34] proposed an adversarial framework utilizing low and high-frequency prior-based discriminators for domain invariant dehazing. While these works highlight information represented within the frequency domain and devise different strategies for exploiting it for restoration tasks, we use an adversarial Fourier network to

leverage its duality property with an image for performing region-sensitive image enhancement. As sRGB images are widely available and used for conducting research for high-level perception tasks, [23] constructed a dataset to demonstrate that low illumination conditions obscure information contained within an image, resulting in a performance drop of SoTA object detectors. Theoretically, the performance can be retained or improved if the image is processed using an ideal enhancement algorithm. However, from our experiments, we demonstrate that current methods result in increased pixelation and noise that adversely affect performance. Hence an ideal image enhancement algorithm is still missing.

2.2 End-to-End Camera ISP

Traditional ISP comprises multiple low-level tasks such as white balancing, demosaicing of CFA data, denoising, high dynamic range compression, black pixel removal, contrast enhancement, tone mapping, super-resolution, etc. The order of application and additional algorithms are unique to sensor manufacturers and inaccessible in most cases with extensive studies being conducted for independently performing these low-level tasks with state-of-the-art (SoTA) performance achieved using CNNs. Furthermore, due to the electronic nature of the camera sensor, it is prone to various noise from various sources such as photon noise, quantization noise, and digital noise [8, 45]. This motivated different works such as [1, 6, 30] to focus on removing noise to improve the signal-to-noise ratio, which has a more prominent effect on images captured in low light conditions [4, 45] due to low pixel intensities. [16] observed superior performance of dehazing algorithm with reduced artifacts when the RAW image is used instead of sRGB image. Encouraged by the success of individual CNNs on low-level tasks, [14] proposed a solution to jointly perform denoising and demosaicing using a residual connection to improve feature flow and better leverage image structure. [29] extended this approach by integrating the task of super-resolution and jointly optimizing the underlying CNN to obtain high-quality RGB images. To further improve the quality of sRGB images [17, 32] proposed a two-stage framework for sequentially restoring and enhancing an image. Lately [10] proposed an end-to-end framework for mapping a RAW demosaiced image captured via smartphone camera into sRGB space while simultaneously enhancing it to match the quality with a DSLR camera. Similar to these approaches, we perform end-to-end RAW-to-sRGB image conversion while removing illumination inconsistencies to obtain a balanced image using a demosaiced image as input and improving performance using frequency priors while achieving real-time performance.

3 Methodology

3.1 Problem Formulation

Functioning of current SoTA image enhancement algorithms is limited to either low illumination or high contrast conditions while being capped by non-linearities arising from camera-ISP. This increases the computational cost of current SoTA, making them unviable for real scenarios wherein such image enhancement mechanisms can improve the performance of different perception tasks. As frequency spectrum is beneficial in ascertaining the limitations of current SoTA algorithms, we integrate such information in the CNN architecture and optimization cycle to ensure textural and structural consistency within enhanced images without noise or unwanted artifacts, thereby providing high dynamic range without relying upon multiple images.

3.2 Network Architecture

Our baseline architecture comprises a two-stage process wherein the first stage enhances illumination, and the second stage removes any artifacts (Fig. 4). We utilize different techniques

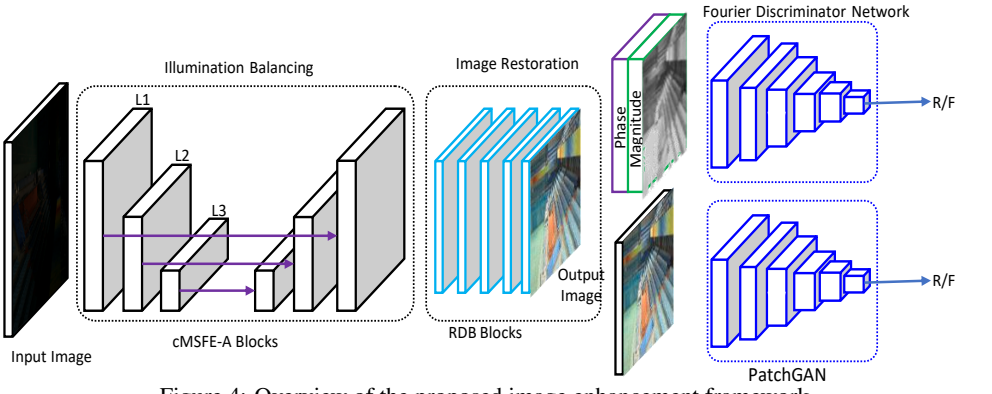

Figure 4: Overview of the proposed image enhancement framework

introduced in the literature to develop efficient modules that provide high feature quality with reduced computational resources.

Illumination Balancing - As different regions can be affected by various illumination sources, to ensure consistent illumination, we use a UNet [31] style encoder-decoder framework that allows convolutional kernels access to complete image at the encoder end. To avoid excessive computations arising from extracting features across multiple scales (1/2, 1/4, 1/8, 1/16, 1/32), inspired by performance gains achieved by increasing receptive field size of convolutional layers, we instead focus on achieving compact multi-scale feature extraction mechanism and thus extract features from 3 scales, i.e., 1/2, 1/8 and 1/32. Using multi-scale convolutions increases computational cost, hence to reduce the computational cost, different techniques such as channel shuffling, [50], squeeze-excitation [9], inception modules [39], 1x1 point convolutions [19] etc., were proposed. In this paper, we combine these techniques to obtain a computationally efficient multi-scale feature extraction mechanism (Fig. 3) and propose a channel split mechanism that divides a feature map across channel dimensions into multiple parts (P) of equal size (implying number of channels (C) should be divisible by a split factor S). These parts are then used as inputs for convolutional kernels of different filter sizes to obtain features across a wider receptive field. To ensure relevant scale-specific features are amplified, we integrate a channel attention mechanism to features from each scale which are subsequently aggregated.

Image Restoration - Upon enhancing illumination within the image, different artifacts and noises can be effectively restored. Furthermore, to ensure the presence of structural and textural details, we concatenate the original input image along with enhanced image and use residual dense blocks (RDBs) [53] (that combine local and global features) to ensure similarity with ground truth. While RDBs are usually used on down-scaled features, we argue that subsequent upsampling of these features would reduce image quality resulting in losing details. Contrarily using them on complete images avoids these issues.

3.3 Fourier Adversarial Network

As there is a perceivable difference between Fourier transforms of low light and corresponding ground truth images (Fig. 2), frequency domain information (extracted using Fast Fourier Transform (FFT)) can be used to improve image quality by incorporating it within the optimization cycle. Furthermore, as low and high frequencies can be used concurrently to capture structural information better, using a complete Fourier spectrum can ensure structural consistency within the enhanced image. While pixel-based losses could be used to ensure FFT of enhanced and ground truth images are similar, they fail to capture inter-pixel

relationships across neighboring pixels. Instead, we propose to use a CNN-based binary loss to determine whether a given image is real or fake based on the Fourier spectrum of its grayscale version. Furthermore, since the magnitude spectrum of an image has higher intensity around zero frequency, we normalize it before concatenating it with an input image that is then passed to the discriminator for binary classification. As we use a Fourier-based adversarial network to identify real/fake images using structural details, we require another adversarial network to ensure equal balance towards textural details. Thus we use commonly used PatchGAN [11] for this purpose.

In summary, the complete framework comprises a two-stage enhancement network that acts as a generator (G) with two discriminators focusing on structural (D1) and textural (D2) details to determine genuinity of a given image. For optimizing the complete framework, a combination of pixel (L1), structural (MS-SSIM [42]), and feature-based (Supervised Contrastive Adversarial Loss) loss functions along with adversarial losses (following LSGAN [26]) are used resulting in the following optimization objective for learnable parameters within generator (θ_G) and discriminators (θ_{D1}, θ_{D2}),

$$\begin{aligned}\min_{\theta_G} \max_{\theta_{D1},\theta_{D2}} \quad & \lambda_{L1}\mathbb{L}_{L1}(\theta_G) + \lambda_{MS-SSIM}\mathbb{L}_{MS-SSIM}(\theta_G) \\ & + \lambda_{SCAL}\mathbb{L}_{SCAL}(\theta_G) + \lambda_{P-ADV}\mathbb{L}_{P-ADV}(\theta_G,\theta_{D1}) + \lambda_{F-ADV}\mathbb{L}_{F-ADV}(\theta_G,\theta_{D2})\end{aligned} \tag{1}$$

Here $\lambda_{L1}, \lambda_{MS-SSIM}, \lambda_{SCAL}$ represent weights for balancing the L1, MS-SSIM, and SCAL losses and are set to 1, 1, and 0.01, whereas $\lambda_{P-ADV}, \lambda_{F-ADV}$ represent the weights for balancing the adversarial losses and are set to 0.5. We refer to the generator trained using this process as AFNet.

4 Experimental Evaluations

4.1 Datasets and Evaluation Metrics

As we analyze the performance of different SoTA image enhancement algorithms along with their application in real perception tasks, we rely upon multiple datasets with different evaluation metrics. Hence we categorize them according to tasks and summarize them as,

Image Enhancement - To examine the performance of SoTA algorithms under diverse illumination conditions, exhaustive experiments are performed using datasets containing both sRGB and RAW images. For sRGB images, we use LOL [44] and SICE [3] datasets wherein the LOL dataset contains 1000, 485, 15 paired training, validation, and test images, whereas the SICE dataset contains 400, 130, 58 paired images captured under different illumination conditions ranging from -3ev to +3ev with increments of 1ev. For the RAW dataset, we utilize the SID-Sony [4] subset having 2421, 276 training and test samples along with the ELD [45] dataset that comprises 384, 96 training and test image pairs captured using different camera sensors. To quantify performance, we use a wide range of metrics covering pixel information (PSNR), structural consistency (SSIM [43]) and Textural Consistency both with (LPIPS [49]) and without (NIQE [28]) reference.

Perception Algorithms - To analyze the impact of variable illumination conditions on common perception tasks such as object detection and semantic segmentation, we choose ExDark [23], JOL [36], COCO [20] and Cityscapes [5] datasets and use mAP and mIOU metrics to quantify performance in dynamic illumination conditions. While Exdark and JOL captures dynamic illumination conditions for object detection, we extend the COCO and Cityscapes datasets to represent night conditions using image translation methods to verify the results across datasets and tasks. Specifically, we improve performance of original Cycle-GAN [55]

Table 1: Quantitative results of ablation studies.

Config.	PSNR / SSIM	GMACs
Single Stage	14.25 / 0.57	0.46
Two Stage w 1x RDB	18.16 / 0.62	0.57
w 3x RDB	19.67 / 0.64	0.92
w 5x RDB	20.48 / 0.69	1.58
w 7x RDB	20.51 / 0.71	2.23
Two Stage w 7x RDB + cMSFE-A	21.45 / 0.78	4.38
+ Patch GAN	21.97 / 0.82	4.38
+ Fourier GAN (RGB)	22.98 / 0.83	4.38
AFNet (+ Fourier GAN (Gray))	23.01 / 0.84	4.38

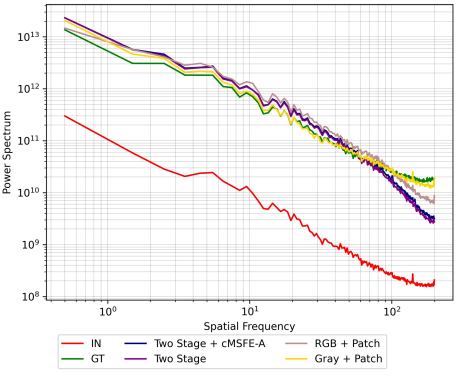

Figure 5: Power Spectral Density (PSD) of images generated using different configurations of proposed framework.

by introducing different techniques which are provided in supplementary. While such techniques could be inversely used to generate well-illuminated images, the computational cost associated with these algorithms to process high-resolution images overshadows their style translation performance.

4.2 Training Mechanism

To enhance sRGB images, we utilized images from the LOL dataset, cropped to 256 x 256 along with augmentation techniques mentioned [35] that ensure the presence of different illumination conditions within training samples. We used an ADAM [13] optimizer with an initial learning rate of 1e-4 for both generator and discriminators. We train the complete framework for 1000 epochs while reducing the learning rate by a factor of 0.5 every 200 epochs and use the model weights that result in minimum validation error across the training process. As we study if camera-ISP could be integrated within the enhancement pipeline, we modify all prior sRGB algorithms to accept 4 channel demosaiced input and use bicubic upsampling to match the resolution of generated images with ground truth. Furthermore, we follow the training process mentioned above without making any modifications to the underlying CNN architecture. For our experiments, we use a system equipped with NVidia 3090 GPU running Pytorch 1.7.

4.3 Ablation Studies

In this section, we examine the effect of different architectural and optimization modifications on network performance in terms of computational cost as well as the quality of the enhanced image. To carry out our examination, we use the LOL dataset as its small size allows us to explore different network variants while minimizing the training time. We first compare the performance of single-stage and two-stage networks that are trained in an end-to-end manner without using pixel, structural and perceptual losses. From performance results summarized in Tab. 1 and PSD in Fig.,5 we observe the two-stage network with 5 RDB blocks to result in peak performance in PSNR and SSIM. We further replace the standard convolutional layers with proposed cMSFE-A layers and observe performance to improve significantly both in terms of quantitative metrics and the PSD curves as well with a large textural component matching the ground truth PSD. Subsequently, we examine the effect of training proposed two-stage enhancement algorithm in GAN framework, specifically focusing on the effect of Fourier spectrum-based discriminator on quality of generated images. Quantitative and qualitative results (In Supplementary) demonstrate that using Fourier

Table 2: Performance Evaluation of SoTA on sRGB images from LOL dataset.

Algorithm	PSNR / SSIM (↑)	NIQE / LPIPS (↓)	GMACs
Input	7.77 / 0.19	5.71 / 0.42	-
DALE	18.55 / 0.73	9.43 / 0.28	211.47
DLN	21.34 / 0.82	3.05 / 0.28	248.02
DSLR	18.22 / 0.62	3.90 / 0.58	18.74
EnlightenGAN	17.48 / 0.65	4.89 / 0.39	61.07
GLAD	19.72 / 0.68	6.80 / 0.40	0.12
KinD	17.65 / 0.77	3.89 / 0.28	14.62
MBLLEN	17.63 / 0.72	3.38 / 0.37	104.76
RetinexNet	16.77 / 0.42	9.73 / 0.47	68.00
URIE	20.10 / 0.72	4.75 / 0.41	14.28
Ours	23.01 / 0.84	3.86 / 0.27	4.38

Table 3: Performance Evaluation of SoTA on RAW images from SID-Sony dataset.

Algorithm	PSNR / SSIM (↑)	NIQE / LPIPS (↓)	GMACs
Rawpy	28.73 / 0.77	4.07 / 0.38	-
EnlightenGAN	24.27 / 0.64	4.68 / 0.53	349.20
RAW2RGB-GAN	23.55 / 0.78	4.00 / 0.71	342.19
KinD	26.91 / 0.73	4.10 / 0.39	196.07
GLAD	27.11 / 0.82	3.86 / 0.39	132.64
SID	28.88 / 0.78	4.39 / 0.43	562.06
TENet	30.17 / 0.83	3.18 / 0.31	1560.14
PyNet	29.01 / 0.79	3.79 / 0.34	2097.03
PyNet-CA	27.24 / 0.74	4.02 / 0.41	2194.14
AWNet	28.09 / 0.76	3.98 / 0.39	460.29
Ours	27.67 / 0.84	3.94 / 0.37	168.08

discriminator improves image generation quality, while the method of extracting Fourier spectrum from a grayscale image or per-channel of RGB image doesn't make a significant difference on performance. From the PSD curves, we can verify that using Fourier adversarial networks indeed improves the performance of enhancement algorithms in the higher frequency spectrum. (We present extended analysis in Supplementary).

4.4 Performance Evaluation with SoTA Algorithms

For comparing enhancement performance on sRGB images, we choose publicly available supervised-learning-based algorithms such as DALE [15], DLN [40], DSLR [18], EnlightenGAN [12], GLAD [41], MBLLEN [25], KinD [51], RetinexNet [44] and URIE [57], whereas for RAW images, we choose Rawpy [1], RAW2RGB-GAN [24], TENet [29], PyNet [10], ELD [45], AWNet [7]. We summarize the qualitative performance on LOL and SID-Sony datasets along with computational requirement in GMAC (Giga- Multiplication and Accumulation Operations)[2] [3] in Tab. 2 and Tab. 3, respectively.

From performance metrics, we can conclude that algorithms comprising multiple subnetworks (GLAD, KinD, DSLR) could provide comparable performance with respect to SoTA (URIE, EnlightenGAN) without consuming excessive computations, with the proposed approach providing new SoTA without excessive computations. In addition, we observe that GAN-based approaches such as EnlightenGAN and AFNet result in improved scores on feature-based metrics such as NIQE and LPIPS, thereby demonstrating GAN-based approaches to generate naturalistic images. In order to examine if these algorithms could be reconfigured to accept demosaiced RAW images and generate enhanced sRGB images, we use EnlightenGAN, KinD, GLAD, and AFNet along with bicubic upsampling mechanism and summarize results in Tab. 3. We observe the performance of these reconfigured algorithms to reach the performance of algorithms that are specifically designed for RAW image enhancement albeit a lower computational requirement.

5 Qualitative Evaluation

We present some visual results demonstrating the effectiveness of proposed approach for qualitative evaluation, with additional examples included in supplementary material. Specifically we show performance comparison with SoTA algorithms on LOL dataset in Fig. 6

[1] https://letmaik.github.io/rawpy/api/index.html
[2] https://github.com/sovrasov/flops-counter.pytorch
[3] Assuming 1 GMACs = 0.5 GFLOPs

along with performance of SoTA object detection algorithms on low light and enhanced images from ExDark dataset [23]. From these results we demonstrate both quantitative and qualitative superiority of the proposed mechanism that aids in performance of SoTA Object detection algorithms.

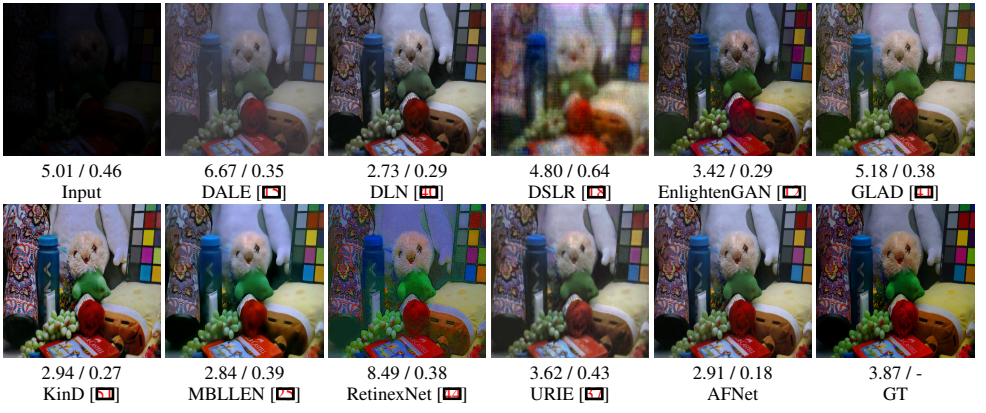

Figure 6: Performance of SoTA algorithms on image from LOL dataset with NIQE / LPIPS score respectively.

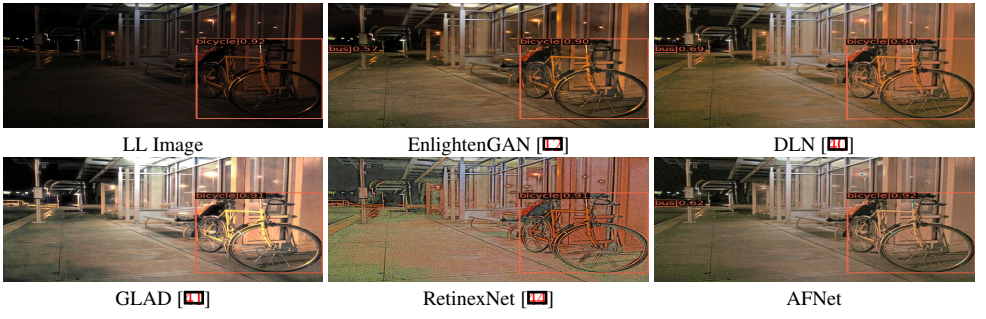

Figure 7: Qualitative Performance of Deformable DETR [56] object detector on Low Light and Enhanced Images following different SoTA.

6 Conclusion

In this paper, we presented the need for balancing illumination present in an image and proposed a fourier adversarial network to ensure presence of structural details within enhanced images. Subsequently we demonstrated the proposed approach to provide SoTA performance while consuming minimum computational resources making it lucrative to be deployed on edge or resource constrained devices. We further demonstrated that Camera-ISP adversely affects the performance of image enhancement algorithms, which can be improved if raw demosaiced images are used as inputs thus the image enhancement algorithm can integrate the functionality of camera-ISP. Finally we demonstrate the varying illumination conditions adversely affect the performance of SoTA object detection and semantic segmentation algorithms which can be improved using image enhancement algorithms.

Acknowledgement This research was supported by KAIST-KU Joint Research Center, KAIST, Korea (N11200035).

References

[1] Abdelrahman Abdelhamed, Mahmoud Afifi, Radu Timofte, and Michael S Brown. Ntire 2020 challenge on real image denoising: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 496–497, 2020.

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[3] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4): 2049–2062, 2018.

[4] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[7] Linhui Dai, Xiaohong Liu, Chengqi Li, and Jun Chen. Awnet: Attentive wavelet network for image isp. *arXiv preprint arXiv:2008.09228*, 2020.

[8] Hao Guan, Liu Liu, Sean Moran, Fenglong Song, and Gregory Slabaugh. Node: Extreme low light raw image denoising using a noise decomposition network. *arXiv preprint arXiv:1909.05249*, 2019.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[10] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. *arXiv preprint arXiv:2002.05509*, 2020.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[12] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *arXiv preprint arXiv:1906.06972*, 2019.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Filippos Kokkinos and Stamatios Lefkimmiatis. Iterative joint image demosaicking and denoising using a residual denoising network. *IEEE Transactions on Image Processing*, 28(8):4177–4188, 2019.

[15] Dokyeong Kwon, Guisik Kim, and Junseok Kwon. Dale: Dark region-aware low-light image enhancement. *arXiv preprint arXiv:2008.12493*, 2020.

[16] Yeejin Lee, Keigo Hirakawa, and Truong Q Nguyen. Joint defogging and demosaicking. *IEEE Transactions on Image Processing*, 26(6):3051–3063, 2016.

[17] Zhetong Liang, Jianrui Cai, Zisheng Cao, and Lei Zhang. Cameranet: A two-stage framework for effective camera isp learning. *arXiv preprint arXiv:1908.01481*, 2019.

[18] Seokjae Lim and Wonjun Kim. Dslr: Deep stacked laplacian restorer for low-light image enhancement. *IEEE Transactions on Multimedia*, 2020.

[19] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[21] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[22] Tong Liu, Zhaowei Chen, Yi Yang, Zehao Wu, and Haowei Li. Lane detection in low-light conditions using an efficient data enhancement: Light conditions style transfer. *arXiv preprint arXiv:2002.01177*, 2020.

[23] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. doi: https://doi.org/10.1016/j.cviu.2018.10.010.

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

[25] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, page 220, 2018.

[26] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[27] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

[28] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[29] Guocheng Qian, Jinjin Gu, Jimmy S Ren, Chao Dong, Furong Zhao, and Juan Lin. Trinity of pixel enhancement: a joint solution for demosaicking, denoising and super-resolution. *arXiv preprint arXiv:1905.02538*, 2019.

[30] Tal Remez, Or Litany, Raja Giryes, and Alex M Bronstein. Deep class-aware image denoising. In *2017 international conference on sampling theory and applications (SampTA)*, pages 138–142. IEEE, 2017.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[32] Eli Schwartz, Raja Giryes, and Alex M Bronstein. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28(2): 912–923, 2018.

[33] Pranjay Shyam, Antyanta Bangunharcana, and Kyung-Soo Kim. Retaining image feature matching performance under low light conditions, 2020.

[34] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Towards domain invariant single image dehazing, 2020.

[35] Pranjay Shyam, Sandeep Singh Sengar, Kuk-Jin Yoon, and Kyung-Soo Kim. Evaluating copy-blend augmentation for low level vision tasks. *arXiv preprint arXiv:2103.05889*, 2021.

[36] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Weakly supervised approach for joint object and lane marking detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2885–2895, October 2021.

[37] Taeyoung Son, Juwon Kang, Namyup Kim, Sunghyun Cho, and Suha Kwak. Urie: Universal image enhancement for visual recognition in the wild. In *ECCV*, 2020.

[38] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion. In *Artificial Intelligence and Machine Learning in Defense Applications*, volume 11169, page 111690A. International Society for Optics and Photonics, 2019.

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[40] Li-Wen Wang, Zhi-Song Liu, Wan-Chi Siu, and Daniel P.K. Lun. Lightening network for low-light image enhancement. *IEEE Transactions on Image Processing*, 2020. doi: 10.1109/TIP.2020.3008396.

[41] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu. Gladnet: Low-light enhancement network with global awareness. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference*, pages 751–755. IEEE, 2018.

[42] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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

[45] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[46] Ke Xu, Xin Yang, Baocai Yin, and Rynson W.H. Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[47] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020.

[48] Cheng Zhang, Qingsen Yan, Yu Zhu, Xianjun Li, Jinqiu Sun, and Yanning Zhang. Attention-based network for low-light image enhancement. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[50] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[51] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 1632–1640, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6889-6. doi: 10.1145/3343031.3350926. URL http://doi.acm.org/10.1145/3343031.3350926.

[52] Yu Zhang, Xiaoguang Di, Bin Zhang, and Chunhui Wang. Self-supervised image enhancement network: Training with low light images only. *arXiv*, pages arXiv–2002, 2020.

[53] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

[54] Yuzhi Zhao, Lai-Man Po, Tiantian Zhang, Zongbang Liao, Xiang Shi, et al. Saliency map-aided generative adversarial network for raw to rgb mapping. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3449–3457. IEEE, 2019.

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[56] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gZ9hCDWe6ke.