Low-Light Image Enhancement for UAVs With Multi-Feature Fusion Deep Neural Networks

Anirudh Singh, Amit Chougule, Pratik Narang, *Senior Member, IEEE*, Vinay Chamola, *Senior Member, IEEE*, and F. Richard Yu, *Fellow, IEEE*

***Abstract*— Object detection in low-light aerial images is a challenging problem due to considerable variation in brightness and varying contrast. Deep learning-based approaches have recently demonstrated great promise in image enhancement. Many existing neural networks used for image quality enhancement first encode the input into low-resolution representations and then decode these representations back to a higher resolution for the contextual information. However, this method leads to the loss of semantic content. Recent research has demonstrated the advantage of maintaining high-resolution information along with lower resolution representations, which maintains image features throughout the network. In this letter, we propose a novel architecture named RNet for low-light image enhancement of aerial images. The proposed network contains multiresolution branches for better understanding of different levels of local and global context through different streams. The performance of RNet is evaluated on a recent synthetic dataset. We also present a comprehensive evaluation with a representative set of state-of-the-art enhancement techniques and neural net architectures.**

***Index Terms*— Deep learning, image enhancement, low-light vision, unmanned aerial vehicle (UAV).**

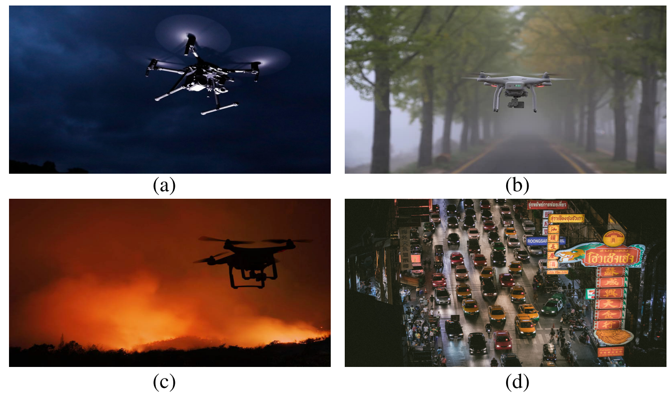

Fig. 1. Utilities of low-light vision capabilities in UAVs. (a) Night Patrol. (b) Operation in foggy environments. (c) Rescue operations. (d) Night time surveillance.

I. Introduction

AERIAL images captured by unmanned aerial vehicle (UAV) are widely used in numerous applications due to their high information content. The weather and lighting circumstances make capture more difficult, resulting in noise, color distortion, or darker images in aerial images, which may reduce the performance of many computer vision applications. The capability to enhance low-light images is crucial for building autonomous vehicles operated under low-light conditions, some of which are shown in Fig. 1. Traditionally, well-lit images have been captured by changing factors, such as the camera aperture, increasing the exposure time, using burst captures, or using artificial lighting. Each of these methods has their drawbacks. For instance, increasing the camera aperture introduces depth of field effects, while flashes and artificial lighting have little to no effect in the aerial context where the lighting conditions are dependent on the weather, which is unalterable.

The loss of contrast and luminance in aerial images directly affects other computer vision tasks, such as object detection and segmentation. There is a need to increase image quality in terms of light or brightness and to reduce noise without compromising the quality. Thus, image enhancement for aerial images has emerged as a promising research area in the recent past. Recently, the use of multiresolution features has been proposed for image enhancement [1], [2]. However, the work has been quite limited. Most of the existing works operate on indoor or outdoor ground images, and their efficacy is not tested on above-ground images.

The contributions of this letter are summarized as follows.

1) We propose a novel deep neural net architecture named RNet for the task of low-light image enhancement for aerial vehicles, which preserves color information in the output images.
2) RNet leverages high-resolution features to extract the rich local semantic information along with

Manuscript received 29 January 2022; revised 21 April 2022; accepted 25 May 2022. Date of publication 8 June 2022; date of current version 19 September 2022. The work of Pratik Narang was supported in part by the National Rural Infrastructure Development Agency (NRIDA) Research Grant funding under the Project titled “AI-enabled drone-based remote health assessment of PMGSY roads” under Grant NRRDA-P017(23)/2021-Dir (P-II) and in part by ARTPARK Student Innovation Grant funding under the Project titled “Enhancing drone-based surveillance in low visibility conditions” under Project UG-07. The work of Vinay Chamola and F. Richard Yu was supported by the Shastri Indo Canadian Institute (SICI) Shastri Institutional Collaborative Research Grant (SICRG) through the Artificial Intelligence Enabled Security Provisioning and Vehicular Vision Innovations for Autonomous Vehicles Project. *(Corresponding authors: Pratik Narang; Vinay Chamola.)*

Anirudh Singh and Pratik Narang are with the Department of Computer Science and Information Systems, Birla Institute of Technology and Science (BITS)-Pilani, Pilani 333031, India (e-mail: f20190107@pilani.bits-pilani.ac.in; pratik.narang@pilani.bits-pilani.ac.in).

Amit Chougule and Vinay Chamola are with the Department of Electrical and Electronics Engineering, and the Anuradha and Prashanth Palakurthi Centre for Artificial Intelligence Research (APPCAIR), Birla Institute of Technology and Science (BITS)-Pilani, Pilani 333031, India (e-mail: amitchougule121@gmail.com; vinay.chamola@pilani.bits-pilani.ac.in).

F. Richard Yu is with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: richard.yu@carleton.ca).

Digital Object Identifier 10.1109/LGRS.2022.3181106

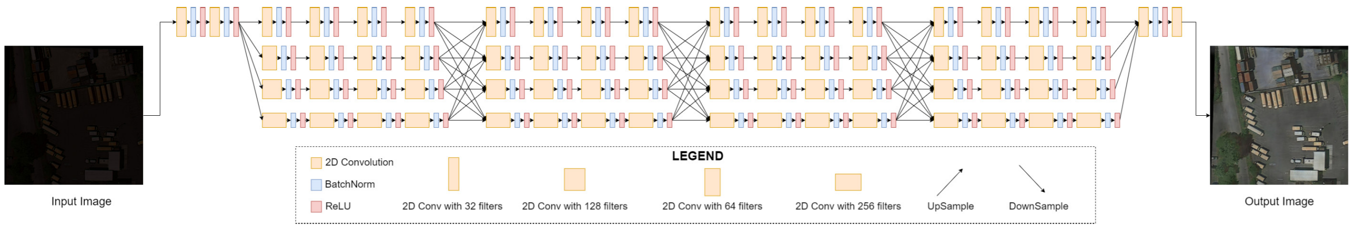

Fig. 2. Proposed network architecture with multiple branches of varying feature resolutions. The highest stream contains features with the highest resolution, which becomes 1/4th, 1/16th, and so on for the lower streams.

low-resolution representations to understand the global context throughout the network. The information is exchanged between the networks at multiple stages to ensure efficient understanding of the scene.

3) The architecture utilizes multiscale feature fusion, which fuses image features at multiple scales in an adaptive manner.
4) The proposed network outperforms existing state-of-the-art methods and traditional network architectures on a synthetic low-light dataset based on a large-scale dataset for instance segmentation in aerial images (iSAID) [3].

II. Related Works

Traditionally, low-light image enhancement has been done through image processing algorithms. Many techniques based on histogram equalization (HE) [4] or gamma correction [5] have been developed. Recently, deep learning techniques have been developed for low-light image enhancement as well. Chen *et al.* [6] worked with raw Bayer arrays to avoid compression losses and use a deep convolutional network to get the well-lit outputs. Zhang *et al.* [7] use attention-based networks to transform low-light images. Lore *et al.* [8] used deep autoencoders, and Arora *et al.* [9] applied global and local context modeling, and Guo *et al.* [10] suggested regularizing the illumination map for maritime images.

Wang *et al.* [11] proposed a technique to improve underexposed images. They use intermediate illumination in the network, which allows the network to learn complicated photographic modifications to convert underexposed input samples into their expert-like retouched outputs. Guo *et al.* [12] proposed a deep curve estimation net (DCE-Net), which formulates light enhancement as a task of image-specific curve estimation. The curve estimation is based on pixel value range, monotonicity, and differentiability. Garg *et al.* [13] developed a strategy for improving aerial instance segmentation through a self-supervised method for the aerial context. Retinex-based methods [14], [15] assume that images can be decomposed into reflectance and illuminance. They decompose and manipulate these two segments to enhance the image samples. Jiang *et al.* [16] and Hua and Xia [17] used generative networks with the retinex theory and propose an approach that does not require paired samples. Chen *et al.* [18] suggested a technique, which learns a photograph enhancer from a group of images with the required qualities, which transforms an input image into an enhanced image with those characteristics.

Few approaches try to incorporate multilevel or multiresolution features as well. Wang *et al.* [1] used multilevel features and used attention to fuse the different feature maps. Zamir *et al.* [2] used spatially precise high-resolution representations as well as low-resolution representations to extract the contextual information.

III. Method

We propose an end-to-end pipeline with multiscale feature fusion based on [21] to enhance low-light images (see Fig. 2). We compare the performance of common architectures to solve the image enhancement problem and also propose a modification of one.

A. Network Architecture

Inspired from [21], we use multiple streams of features connected in parallel to extract both high- and low-frequency details. The higher resolution streams have lower width (a lesser number of convolutional filters) and the number of filters increases as the resolution goes down. This is done to equally distribute the computations within different streams. This approach also ensures that all the streams have a similar number of parameters and learning capability.

The input image is passed through two convolutional layers. The feature map is then downsampled to four different levels and sent to different resolution branches. Each resolution branch contains four stages, where each stage itself contains four convolutional blocks. Each block consists of a 2-D convolution with 3×3 kernels, followed by Batchnorm and a ReLU activation. To connect the different resolution features between different streams, we perform downsampling and upsampling operations on the outputs of each stage. The downsampling is done via strided convolutions with 3×3 kernels and upsampling is done using the nearest neighbor algorithm.

To ensure that the proposed model does not have an unfair advantage due to the higher number of trainable parameters, we train two networks: RNet and RNet-lite. They have the same overall structure and only differ in the number of convolutional filters, as detailed in Table I.

B. Multiscale Feature Fusion

Let $\mathcal{N}_{sr}$ denote the subnetwork in the sth stage and r be its resolution index. A subnetwork with resolution index r outputs a feature map of resolution $(1/2^{(r-1)})$th the input resolution. $\{\mathbf{R}_r^i, r = 1, 2, 3, 4\}$ denotes the input representations to the rth resolution index, and $\{\mathbf{R}_r^o, r = 1, 2, 3, 4\}$ denotes the output from the rth resolution index. Each of the input representation are transformed (either upsampled or downsampled by the

TABLE I
ARCHITECTURE OF THE TWO PROPOSED NETWORKS

RNet-lite	RNet
Subnetwork 1	
ConvBlock-(N20,K3,S1)	ConvBlock-(N32,K3,S1)
ConvBlock-(N20,K3,S1)	ConvBlock-(N32,K3,S1)
ConvBlock-(N20,K3,S1)	ConvBlock-(N32,K3,S1)
ConvBlock-(N20,K3,S1)	ConvBlock-(N32,K3,S1)
Subnetwork 2	
ConvBlock-(N40,K3,S1)	ConvBlock-(N64,K3,S1)
ConvBlock-(N40,K3,S1)	ConvBlock-(N64,K3,S1)
ConvBlock-(N40,K3,S1)	ConvBlock-(N64,K3,S1)
ConvBlock-(N40,K3,S1)	ConvBlock-(N64,K3,S1)
Subnetwork 3	
ConvBlock-(N80,K3,S1)	ConvBlock-(N128,K3,S1)
ConvBlock-(N80,K3,S1)	ConvBlock-(N128,K3,S1)
ConvBlock-(N80,K3,S1)	ConvBlock-(N128,K3,S1)
ConvBlock-(N80,K3,S1)	ConvBlock-(N128,K3,S1)
Subnetwork 4	
ConvBlock-(N160,K3,S1)	ConvBlock-(N256,K3,S1)
ConvBlock-(N160,K3,S1)	ConvBlock-(N256,K3,S1)
ConvBlock-(N160,K3,S1)	ConvBlock-(N256,K3,S1)
ConvBlock-(N160,K3,S1)	ConvBlock-(N256,K3,S1)

N, K, and S denote the number of convolutional filters, Kernel size and stride for the 2D convolution for each ConvBlock, respectively.

scale $2^{r_2-r_1}$, depending on r_1 and r_2). Let $f_{r_1r_2}()$ denote the transformation function, and then, $f_{1,2}(\mathbf{R}_1^i)$, for instance, represents the $2\times$ upsampled representation of $\mathbf{R}_1^i$.

The output is derived from the transformed input representations as

$$\mathbf{R}_r^o = \sum_{k=1}^{k=4} f_{kr}(\mathbf{R}_k^i). \tag{1}$$

Each of $\mathbf{R}_r^o$ is then passed through four conv-blocks containing residual connections. The subnetwork composition in the proposed network can be given as

$$\begin{aligned} &\mathcal{N}_{11} \rightarrow \mathcal{N}_{21} \rightarrow \mathcal{N}_{31} \rightarrow \mathcal{N}_{41} \\ &\mathcal{N}_{12} \rightarrow \mathcal{N}_{22} \rightarrow \mathcal{N}_{32} \rightarrow \mathcal{N}_{42} \\ &\mathcal{N}_{13} \rightarrow \mathcal{N}_{23} \rightarrow \mathcal{N}_{33} \rightarrow \mathcal{N}_{43} \\ &\mathcal{N}_{14} \rightarrow \mathcal{N}_{24} \rightarrow \mathcal{N}_{34} \rightarrow \mathcal{N}_{44}. \end{aligned} \tag{2}$$

Note that the features are shared among different streams as well, as shown in Fig. 2.

IV. EXPERIMENTS

A. Training

Due to the lack of paired low-light aerial images, we use a subset of the synthetic dataset from [13]. It contains images from [3], and the corresponding low-light samples are generated using CycleGAN. We train all the networks with AdamW optimizer and one-cycle policy for cyclical learning rates. Most of the models were trained for 110 epochs on the dataset except RNet and UNet-v1 that were trained for an additional ten epochs until convergence. The networks were trained with 9895 and 402 paired samples in the training set and validation set, respectively. The images in the dataset are 800×800 px. During training, we resize them to 384×384 px ($384 = 1.5 \times 256$) and then crop a random 256×256 px section. We also apply random horizontal and vertical flipping.

All the networks are trained on a NVIDIA titan v GPU with 12 GB of memory and a batch size of 4. We use mse as the loss function.

B. Evaluation Metrics

The evaluation metrics used are peak signal to noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS) [23] to measure perceptual similarity and a no-reference image quality assessment metric Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [24] PSNR is defined as: $\mathrm{PSNR}(x, y) = 10\log_{10}(\mathrm{MAX}^2/\mathrm{mse}(x, y))$, where $\mathrm{MAX} = 1$ (since we scale the pixel values to $[0, 1]$ before using). $\mathrm{mse}(x, y)$ is the mean squared error between images x and y. Also,

$\mathrm{SSIM}(x, y) = ((2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)/(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2))$ is calculated for each of the three input channels and averaged to get the final value. Here, σ_x and σ_y denotes the standard deviation of the pixel values of x and y. μ_x, μ_y are the means and σ_{xy} is the covariance between x and y. The LPIPS distance is used to measure the perceptual similarity between the enhanced and the target images. It is computed using the activations of the AlexNet network, trained on the Imagenet dataset. BRISQUE is a no-reference metric and predicts the human judgment of quality without reference images. It uses image statistics of locally normalized luminance coefficients to quantify possible losses of naturalness in the image.

V. RESULTS

A. Comparison With State of the Art

We compare the result with traditional Image enhancement techniques: HE and contrast limited adaptive histogram equalization (CLAHE). The input image in red green blue (RGB) is converted to hue saturation value (HSV), so the illuminance channel can be equalized and then converted back to RGB. The output of these techniques suffers from color distortion and noise. We also compare the output with Retinex-based techniques, namely, Low-light Image Enhancement via Illumination Map Estimation (LIME) [19] and Dual Illumination Estimation for Robust Exposure Correction (DUAL) [20]. LIME estimates the illumination map by finding the maximum of R, G, and B values for each pixel in the input image. The illumination map is then refined by imposing a structural prior on it. This helps improve the brightness and contrast of the image, but the output again suffers from color distortion. Fig. 3 presents output comparison for various approaches.

We also compare the performance of different architectures over the same approach. UNet, although mostly used for segmentation tasks, performs well in image enhancement, as shown in [6]. We use different scale UNets and the HRnet as baselines to compare our network with. The training performance with epochs is shown in Fig. 4.

Both the proposed networks, RNet and the smaller RNet-lite, outperform similar networks and other image enhancement techniques in almost all of the metrics used, as shown in Table II. Fig. 5 represents number of parameters versus validation PSNR graph. The exchange of feature representations among different branches provides the benefits of group convolutions. The multiscale features simulate the effect of dilated convolutions and hence allow the kernels to extract

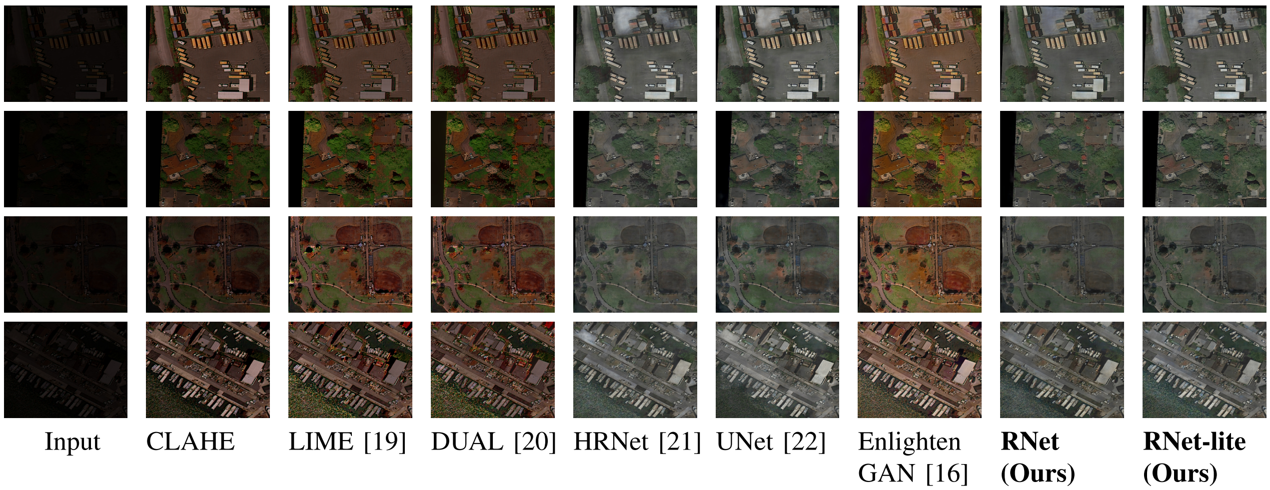

Fig. 3. Output comparison of different approaches. The images generated with RNet are more visually pleasing than all other networks. The multilevel streams in our network help maintain a consistent level of brightness in the image, while the output of traditional networks suffers from overenhanced and underenhanced patches.

TABLE II
COMPARISON OF DIFFERENT TECHNIQUES

Network or technique	PSNR	SSIM	LPIPS	BRISQUE
LIME [10]	16.36	0.60	0.317	47.8
DUAL [20]	16.47	0.60	0.316	47.6
HE [4]	13.32	0.44	0.434	52.1
CLAHE	17.59	0.62	0.238	43.8
UNet-v1	26.89	0.89	0.071	**42.9**
UNet-v2 [22]	25.58	0.86	0.092	44.6
HRnet [21]	27.24	0.90	0.061	44.2
RNet-lite(Ours)	27.81	**0.91**	0.056	**42.9**
RNet(Ours)	**28.09**	**0.91**	**0.052**	43.5

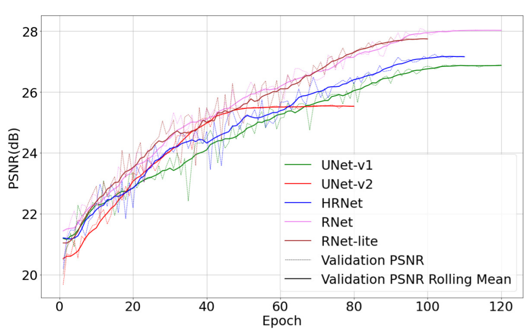

Fig. 4. Validation PSNR with training epochs. UNet-v2 is UNet [22], and UNet-v1 has the same structure as UNet-v2, but with double the number of filters in each convolutional layer. RNet and RNet-lite are the proposed networks with a different number of parameters. UNet-v2 saturated before other networks, and hence, the training was paused after 80 epochs.

the contextual information over a larger scene, not just the neighboring pixels.

We also include some recent approaches such as EnlightenGAN [16] in this section (see Fig. 3). Since these have different training domains, we did not report the metrics for EnlightenGAN but present the results for qualitative analysis.

To analyze the real-world performance of the network, we also present the images enhanced by RNet in Fig. 6 on night-time drone and satellite images.

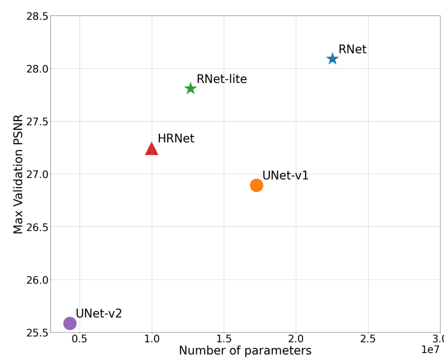

Fig. 5. Number of parameters versus validation PSNR.

TABLE III
EFFECT OF PARALLEL STREAMS AND MULTIRESOLUTION FEATURES

Network	PSNR	SSIM	LPIPS	MSE
RNet	**28.09**	**0.91**	**0.052**	**0.0026**
RNet w/o parallel streams	24.55	0.82	0.105	0.0050
RNet w/o downsampling in - parallel streams	27.6	0.90	0.059	0.0029

VI. ABLATION STUDY

We perform ablation studies to check the effect of both parallel feature streams and multiresolution features. All the networks compared here were built to require similar amount of compute. Compute requirements are measured through the total number of multiply–accumulate (MAC) operations each network requires.

The network without the parallel streams has a single stream of the same scale as the highest stream in RNet. To increase the number of parameters, we increased the number of kernels in each convolution. The number of convolutional kernels was also changed similarly in the other ablated network with parallel streams but a single resolution of features. As shown in Table III, both the ablated networks perform worse than the proposed network (RNet) in all the metrics used. This indicates

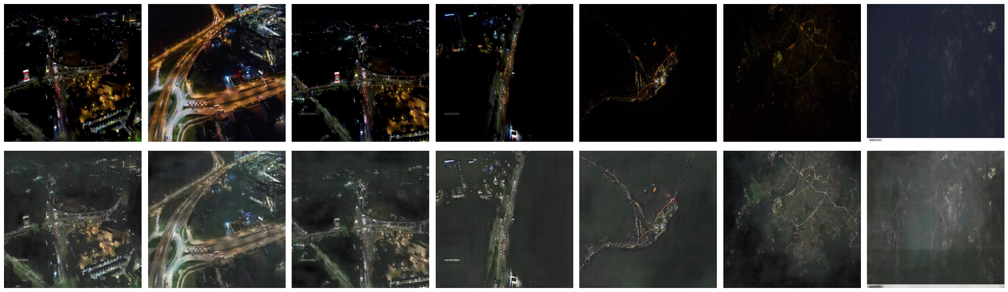

Fig. 6. Performance on real-world images. (Top row) Real-world low-light images. (Bottom row) Images enhanced using RNet. Although the enhanced images contain more information than the original dark images, they are not very visually appealing. This might be because the network was trained on a different dataset than the real-world images captured in different environments at different heights and angles.

that both parallel features streams and multiresolution features positively impact the network performance. The experiments also indicate that the impact of parallel streams is larger than multiresolution features, causing a larger improvement in scores.

VII. Conclusion

In this letter, we present a deep neural network for low-light image enhancement for UAVs. The proposed RNet utilizes multiscale feature fusion by leveraging high-resolution features to extract the rich local semantic information along with low-resolution representations to understand the global context. By incorporating multiple high-resolution streams, the proposed network is able to outperform other deep learning-based approaches as well as conventional enhancement techniques.

References

[1] L. Wang, G. Fu, Z. Jiang, G. Ju, and A. Men, "Low-light image enhancement with attention and multi-level feature fusion," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 276–281.

[2] S. W. Zamir *et al.*, "Learning enriched features for real image restoration and enhancement," in *Computer Vision—ECCV 2020*. Glasgow, U.K.: Springer, Aug. 2020, pp. 492–511.

[3] S. Waqas Zamir *et al.*, "isaid: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 28–37.

[4] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Comput. Vis., Graph., Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.

[5] S. Rahman, M. M. Rahman, M. Abdullah-Al-Wadud, G. D. Al-Quaderi, and M. Shoyaib, "An adaptive gamma correction for image enhancement," *EURASIP J. Image Video Process.*, vol. 2016, no. 1, pp. 1–13, Dec. 2016.

[6] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3291–3300.

[7] C. Zhang, Q. Yan, Y. Zhu, X. Li, J. Sun, and Y. Zhang, "Attention-based network for low-light image enhancement," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[8] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.

[9] A. Arora *et al.*, "Low light image enhancement via global and local context modeling," 2021, *arXiv:2101.00850*.

[10] Y. Guo, Y. Lu, R. W. Liu, M. Yang, and K. T. Chui, "Low-light maritime image enhancement with regularized illumination optimization and deep noise suppression," 2020, *arXiv:2008.03765*.

[11] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6849–6857.

[12] C. Guo *et al.*, "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1780–1789.

[13] P. Garg, M. Mandal, and P. Narang, "Improving aerial instance segmentation in the dark with self-supervised low light enhancement (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 18, pp. 15781–15782.

[14] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–12.

[15] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1632–1640.

[16] Y. Jiang *et al.*, "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.

[17] W. Hua and Y. Xia, "Low-light image enhancement based on joint generative adversarial network and image quality assessment," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2018, pp. 1–6.

[18] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6306–6314.

[19] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.

[20] Q. Zhang, Y. Nie, and W. Zheng, "Dual illumination estimation for robust exposure correction," *Comput. Graph. Forum*, vol. 38, no. 7, pp. 243–252, Oct. 2019.

[21] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Germany: Springer, 2015, pp. 234–241.

[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. CVPR*, Jun. 2018, pp. 586–595.

[24] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.