Thermal-Aware Low-Light Image Enhancement: A Real-World Benchmark and a New Light-Weight Model

Zhen Wang[1*], Yaozu Wu[2,4*], Dongyuan Li[3], Shiyin Tan[1], Zhishuai Yin[2,4†]

[1]Institute of Science Tokyo, Tokyo, Japan
[2]Wuhan University of Technology, Wuhan, China
[3]The University of Tokyo, Tokyo, Japan
[4]Xianhu Laboratory of the Advanced Energy Science and Technology, Foshan, China
zhenwangrs@gmail.com, yaozuwu279@gmail.com, lidy@csis.u-tokyo.ac.jp
tanshiyin@lr.pi.titech.ac.jp, zyin@whut.edu.cn

Abstract

Enhancing images captured under low-light conditions has been a topic of research for several years. Nonetheless, existing image restoration techniques mainly concentrate on reconstructing images from RGB data, often neglecting the possibility of utilizing additional modalities. With the progress in handheld technology, capturing thermal maps with mobile devices has become straightforward. Investigating the integration of thermal data into image restoration presents a valuable research opportunity. Therefore, in this paper, we propose a multimodal low-light image enhancement task based on thermal information and establish a dataset named **TLIE** (**T**hermal-aware **L**ow-light **I**mage **E**nhancement), consisting of 1,113 samples. Each sample in our dataset includes a low-light image, a normal-light image, and the corresponding thermal map. Additionally, based on the TLIE dataset, we develop a multimodal approach that simultaneously processes input images and thermal map data to produce the predicted normal-light images. We compare our method with previous unimodal and multimodal state-of-the-art LIE methods, and the experimental results and detailed ablation studies prove the effectiveness of our method.

Introduction

Low-light image enhancement (LIE) is a critical and challenging task in computer vision. It refers to the process of converting dimly lit photos captured in low-light settings into images with normal lighting. This process is crucial in numerous applications, including the improvement of nighttime video frames and the capturing of photographs in low-light situations (Wei et al. 2018). Several methods have been proposed for low-light image enhancement, and in recent years, the progress in deep learning technology has led to the extensive application of neural network models in LIE tasks. The main datasets currently in use include LOL (v1 (Wei et al. 2018) and v2 (Yang et al. 2021)), SID (Chen et al. 2018), SMID (Chen et al. 2019), SDSD (Wang et al. 2021), LIME (Guo, Li, and Ling 2016), and DICM (Lee, Lee, and Kim 2013). Notably, LOLv1 (Wei et al. 2018) is the first to introduce a large-scale dataset designed specifically for the LIE task, incorporating Retinex theory (Land 1977) into LIE for the first time. Following this, other researchers have introduced a variety of methods for LIE. Some utilize Retinex-based techniques (Zhang et al. 2021; Zhang, Zhang, and Guo 2019), which concentrate on distinguishing the illumination and reflectance elements of an image to improve its visual quality. Others have examined the effectiveness of encoder-decoder architectures (Jin et al. 2023; Xu et al. 2022), paired with attention mechanisms (Zamir et al. 2022; Cai et al. 2023), where attention-based strategies are crucial in making the model focus on significant image regions.

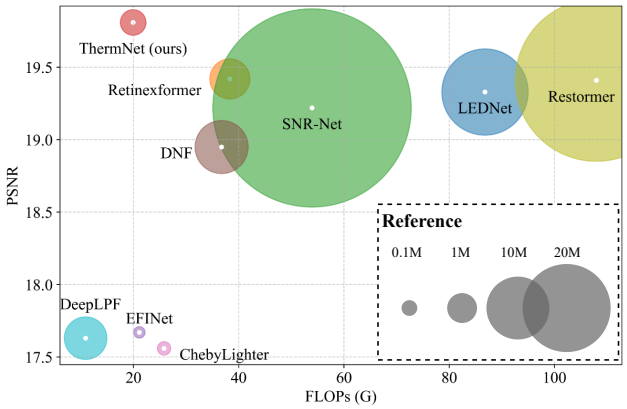

Figure 1: PSNR vs. FLOPs, size to parameters. The size of blobs refers to the account of model parameters (million).

Although there have been recent advancements in the LIE task, there is still significant potential for improvement, particularly with regard to more varied modalities. Nowadays, contemporary devices like drones and self-driving vehicles are outfitted with more sensors compared to ten years ago. Investigating the use of these sensors to improve the quality of photos taken in low-light environments is a very valuable area of research. For example, modern drones now typically come with an infrared camera, which facilitates the effortless acquisition of thermally enhanced images alongside regular ones, especially in the dark night. Compared to other

*These authors contributed equally.
†Corresponding author.

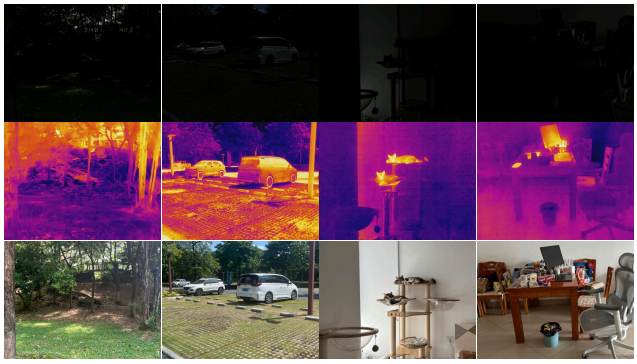

Figure 2: Four samples from TLIE dataset. First row shows low-light RGB images, second row shows the thermal map (colder colors indicate a lower temperature), and third row shows the normal-light images.

non-RGB sensors like LiDAR (Song, Lichtenberg, and Xiao 2015; Silberman et al. 2012; Wang et al. 2024; Wang, Li, and Jiang 2024) or event camera (Jiang et al. 2023), the infrared camera is typically much cheaper, easier to use and less affected by the environment, such as rain or snow. Therefore, it is meaningful to explore ways to incorporate thermal data into the LIE task to achieve multimodal low-light image enhancement (MLIE) (Jiang et al. 2023).

Given the absence of corresponding publicly available datasets, to facilitate research in this field, we use an infrared camera to collect and create a dataset called **TLIE** (**T**hermal-aware **L**ow-light **I**mage **E**nhancement), which comprises 1,113 sets of images, including low-light images, normal-light images, and corresponding thermal maps. Moreover, this study introduces a light-weight model named **ThermNet** designed to effectively incorporate thermal data into the LIE process, which utilizes three techniques to leverage thermal information. The first approach combines SNR maps derived from low-light and thermal data to create an SNR map that more closely resembles the normal-light SNR map, thereby offering supplementary information for restoration. The second technique merges Retinex (Land 1977) theory with multimodality, transitioning the formerly unimodal Retinex algorithm into a multimodal format, resulting in improved and more resilient performance. Moreover, an enhancement sub-network harnessing thermal data is built to further refine the predicted outcomes from the prior stage, leading to more precise results.

Utilizing TLIE and ThermNet, we perform experiments and compare our method against previous state-of-the-art methods in the LIE task, including both unimodal and multimodal methods. The experimental findings show that our method exceeds the accuracy of prior methods while having much less parameters and higher computational efficiency. Moreover, we undertake thorough ablation studies, encompassing both quantitative and qualitative analyses, to demonstrate its effectiveness and flexibility. Our contributions can be summarized as follows:

- We develop an first-of-its-kind dataset for real-world LIE task that incorporates thermal data.
- We introduce an innovative approach called ThermNet, designed to incorporate thermal data into LIE models, thereby enhancing the accuracy of image restoration.
- We showcase the efficiency and versatility of ThermNet through extensive experiments and ablation studies.

Related Work

Recent progress in low-light image enhancement has generated substantial interest in the creation of datasets that truly represent the difficulties encountered in dark conditions. One of the first datasets in this field is the LOL (Wei et al. 2018; Yang et al. 2021) dataset, which covers a wide array of genuine low-light conditions, aiding in the development and assessment of models designed to enhance low-light images. The SID (Chen et al. 2018) and SMID (Chen et al. 2019) datasets are obtained using cameras with short and long exposure times to capture low-light and normal-light images individually. The SDSD (Wang et al. 2021) dataset is created using a camera equipped with an ND filter for both indoor and outdoor settings. Those datasets mainly concentrate on reconstructing normal-light images from provided low-light RGB images. Currently, there are also some datasets that are aimed at the MLIE task, where most of them focus on event images acquired by Event Camera (Jiang et al. 2023; Liang et al. 2024). Compared to event cameras, infrared cameras have several distinct advantages. Firstly, the price of infrared cameras is typically only one-tenth of that of event cameras. Secondly, images obtained from infrared cameras have a higher readability. Additionally, infrared cameras are more user-friendly and easier to operate.

Thermal Datasets. Thermal imaging has seen significant advancements in recent years, with applications spanning various fields such as medical diagnostics, environmental monitoring, and industrial inspection. For instance, the FLIR Thermal Dataset (FLIR 2022) provides a comprehensive collection of thermal maps captured under different conditions, aiding in the development of object detection and classification in thermal spectra. Similarly, the KAIST Multispectral Pedestrian Dataset (Choi et al. 2018) combines thermal and visible spectrum images to enhance pedestrian detection algorithms. LSOTB-TIR Dataset (Liu et al. 2020) uses the thermal map to improve object tracking ability. These datasets, among others, have been instrumental in advancing the state-of-the-art in thermal imaging research, providing benchmarks for evaluating new algorithms and fostering innovation across various applications. Because infrared thermal maps are not affected by ambient lighting, the methods based on them have stronger robustness. Some methods adopt active learning to select the most uncertainty samples among multiple datasets (Li et al. 2024). However, currently there is only one LIE dataset based on thermal maps in TGLLE-Net (Cao et al. 2023), and it is a synthetic dataset, which limits its practical utility. Our TLIE dataset is the first real-world thermal-aware LIE dataset that can be used to promote research in this field.

Dataset

Data Collection. To construct TLIE, we start by capturing thermal maps as well as visible-light images. In this paper, we primarily utilize a portable infrared thermal capture device named K20 provided by Hikvision [1] to record the thermal map and use the rear camera of the iPhone to capture the RGB image. K20 has 256×192 pixel infrared detector and covers 49,152 temperature measurement points, which enables it to record high-quality thermal maps. We use a camera rig to mount the K20 and the iPhone parallel to each other. To capture images in low-light environments, consistent with previous studies (Wei et al. 2018; Chen et al. 2018, 2019), we modify the exposure duration and the camera's ISO settings via the iOS API to reduce brightness and mimic low-light scenarios. For each setup, our process of data gathering includes: (1) secure the camera rig in a fixed position; (2) take a normal-light photograph using the default settings on an iPhone; (3) modify the camera settings and take a photo in low-light conditions; (4) obtain an infrared thermal map with the K20.

Quality Control. To guarantee that each image carries significant thermal information, we made sure that each captured scene included various types of objects, such as desks, walls, and computers. To ensure dataset balance, we recorded scenes encompassing both indoor and outdoor settings, with their proportions being nearly equal. Furthermore, to enhance data diversity, we took only a single image per scene and randomly varied the exposure time and ISO, thus reducing the likelihood of the model leveraging patterns. Additionally, our scenes are captured at different times throughout the day. Once the data have been collected, due to the variance in the field of view between the iPhone and the K20 lenses, we conduct further processing. By cropping and transforming the images, we align the thermal map with the RGB image. Finally, we uniformly resize the resolution of all low-light images, normal-light images, and thermal maps to 640×480. Table 1 presents a comparative analysis of our TLIE against prior LIE datasets, such as LOLv1 (Wei et al. 2018), LOLv2 (Yang et al. 2021), SID (Chen et al. 2018), SDSD (Wang et al. 2021), ELIE (Jiang et al. 2023), SDE (Liang et al. 2024), and TGLLE-Net (Cao et al. 2023). It demonstrates that TLIE is unique as the only real-world dataset for MLIE that includes supplementary thermal data.

Dataset	# Sample	Multimodal	Type	Source	Released
LOLv1	500	None	Image	Real	✔
LOLv2-syn	1,000	None	Image	Synth	✔
SID	2,697	None	Image	Real	✔
SDSD	150	None	Video	Real	✔
ELIE	204	Event	Video	Real	✔
SDE	91	Event	Video	Real	✔
TGLLE-Net	820	Thermal	Image	Synth	✗
TLIE (ours)	1,113	Thermal	Image	Real	✔

Table 1: Comparison of TLIE with different LIE datasets. "# Sample" represents the number of videos.

Methodology

Overview

ThermNet is a dual-phase neural network, firstly employing a Retinex sub-network to generate an initial outcome (as shown in Figure 3). Subsequently, an encoder-decoder refinement sub-network is used to encode, then decode, and progressively up-sample the hidden states layer by layer. Finally, an RGB prediction layer is applied to estimate the color under normal light conditions and then averaged with the preliminary result from the first phase as the final result.

Specifically, in ThermNet, we introduce three additional components to incorporate thermal information into low-light image restoration. First, we propose a *Multimodal SNR Alignment Module* (**MSAM**) to combine the SNR maps from low-light and thermal maps to obtain a SNR map that is closer to that of the normal-light image, therefore providing richer information for image restoration.

Secondarily, prior to feeding the original low-light image directly into the image encoder, we devise a *Multimodal Thermal-based Retinex Network* (**MTRN**) to integrate thermal data with the established Retinex theory, thereby creating a multimodal Retinex method. This technique produces an initial prediction of the image under normal lighting conditions and achieves enhanced outcomes by incorporating extra thermal modal data.

Finally, rather than exclusively depending on simple encoder-decoder layers, we propose the *Thermal-guided Refinement Network* (**TGRN**) to enhance the output from MTRN. Within TGRN, prior to each decoding layer, our custom *Multi-scale Thermal Fusion Layer* (**MTFL**) is employed to integrate the thermal map with the hidden features of the images in each decoder layer. This incremental refinement of decoding features boosts the precision of the results.

Multimodal SNR Alignment Module

SNR map has been used in many previous LIE methods (Xu et al. 2022) to improve the model performance. However, previous LIE methods based on a single RGB image have significant drawbacks. As shown in Figure 3, the SNR map calculated from the low-light image contains many white noise points, especially in very dark areas. This is because SNR calculation heavily relies on RGB differences, while the differences in dark areas are very small. Compared to RGB images, thermal maps, although lacking rich colors, can still capture object contours even in very dark areas, as also shown in the pipeline. The SNR map of thermal maps, compared to that of low-light images, however, lacks certain details. Therefore, we propose a Multimodal SNR Alignment Module (MSAM) that fuses the SNR maps of low-light images and thermal maps to obtain an SNR map closer to that of normal-light images. This fused map is then concatenated with intermediate results to better enhance the image restoration effect. The details of MSAM are shown in Figure 3. In MSAM, during the training phase, we first calcu-

[1] https://www.hikvision.com/

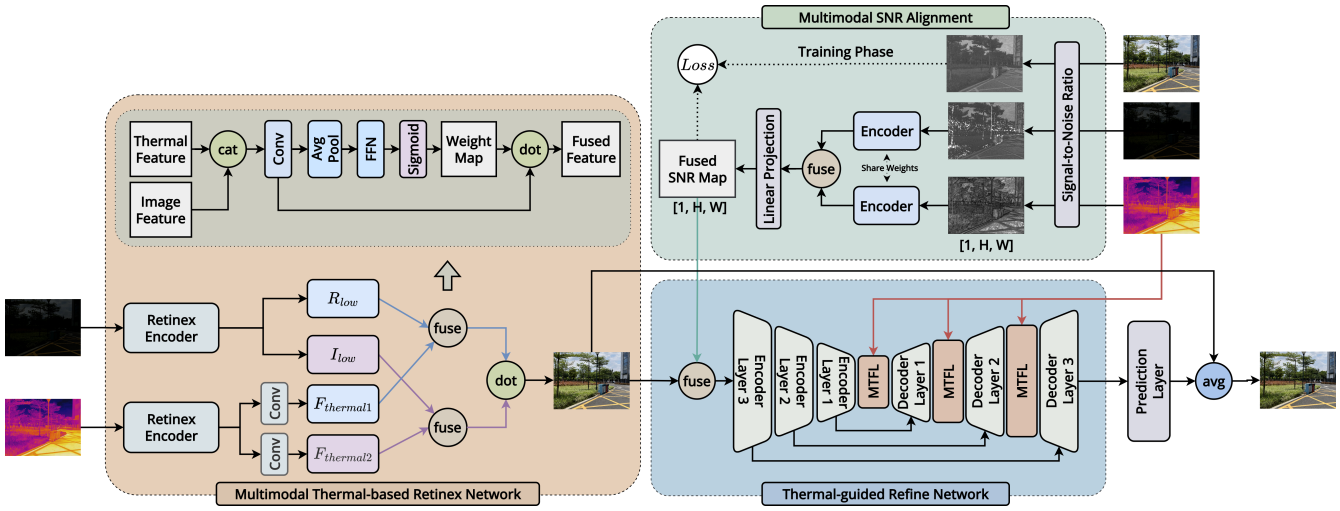

Figure 3: Overall pipeline of our proposed ThermNet. "MTFL" means Multi-scale Thermal Fusion Layer, "cat" means concatenate, "dot" means dot production, "avg" means average. "R" and "I" represent reflectance and illumination, respectively.

late three kinds of SNR map for different types of images, and then we use an encoder to encode a low-light SNR map and a thermal SNR map separately and fuse those two features together with the same way in MTRN. Finally, we apply a linear projection layer to project the fused feature to $\mathbf{F}_{\text{SNR}}$ which has the same shape of normal-light SNR map, and calculate the Mean Squared Error (MSE) Loss $\mathcal{L}_{\text{SNR}}$ between them. During testing phase, only the predicted $\mathbf{F}_{\text{SNR}}$ will be used in following procedure.

Multimodal Thermal-based Retinex Network

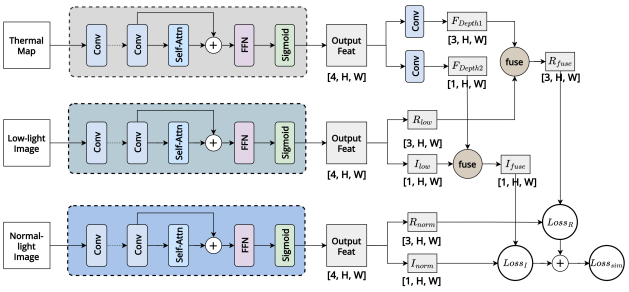

Figure 4: Detail structure of Retinex Encoder and the calculation of normal-light image similarity loss.

The Retinex theory is a conceptual model in computational vision that elucidates how human vision perceives color and brightness in different lighting contexts. It proposes that the human visual system interprets color and brightness by referencing the lightness of objects relative to their surroundings, instead of depending solely on the absolute intensity of light they receive. According to the Retinex theory, an object maintains consistent color regardless of variations in lighting, and the observed images can be separated into two parts: reflectance $\mathbf{R}$ and illumination $\mathbf{I}$. Let S represent the source image, then it can be denoted by:

$$\mathbf{S} = \mathbf{R} \cdot \mathbf{I}, \tag{1}$$

where $\cdot$ means the element-wise dot-production of them.

Retinex finds extensive application in numerous LIE techniques, with Retinex-Net (Land 1977) pioneering the integration of deep learning and Retinex theory. Retinex-Net disassembles $\mathbf{S}$ into $\mathbf{R}$ and $\mathbf{I}$ components through convolutional networks. Building on their research, we are the first in developing a multimodal Retinex approach that concurrently decomposes both low-light images and thermal maps. We then employ a sub-network to independently merge the two variations of $\mathbf{R}$ and $\mathbf{I}$, and ultimately perform element-wise multiplication on the fused $\mathbf{R}$ and $\mathbf{I}$ to generate the predicted normal-light image $\mathbf{S}_{\text{pred}}$.

As illustrated in the left section of Figure 3, within our MTRN, we initially separate $\mathbf{S}_{\text{low}}$ and $\mathbf{S}_{\text{thermal}}$ into $\mathbf{R}_{\text{low}}$ and $\mathbf{I}_{\text{low}}$, and $\mathbf{F}_{\text{thermal1}}$ and $\mathbf{F}_{\text{thermal2}}$, respectively. The specifics of this separation process are displayed in Figure 4. First, we employ our custom-designed Retinex Encoder to separate $\mathbf{S}_{\text{low}}$. It starts by several convolutional layers to the input image, followed by a self-attention layer to give the features a global perspective. Then, a feed-forward neural network (FFN) layer and a sigmoid layer are used to compress the output feature $\mathbf{F}_{\text{RE}} \in \mathbb{R}^{4\times H\times W}$ to 0 and 1. After feature encoding, the first three channels of $\mathbf{F}_{\text{RE}}$ are used as $\mathbf{R} \in \mathbb{R}^{3\times H\times W}$ and the last channel is used as $\mathbf{I} \in \mathbb{R}^{1\times H\times W}$.

To incorporate thermal information into both $\mathbf{R}_{\text{low}}$ and $\mathbf{I}_{\text{low}}$, we utilize an additional Retinex Encoder to encode the thermal map into $\mathbf{F}_{\text{thermal}} \in \mathbb{R}^{4\times H\times W}$. Subsequently, we employ two distinct convolutional layers to compress $\mathbf{F}_{\text{thermal}}$ independently into $\mathbf{F}_{\text{thermal1}} \in \mathbb{R}^{3\times H\times W}$ and $\mathbf{F}_{\text{thermal2}} \in \mathbb{R}^{1\times H\times W}$. We then merge the combinations of ($\mathbf{R}_{\text{low}}$, $\mathbf{F}_{\text{thermal1}}$) and ($\mathbf{I}_{\text{low}}$, $\mathbf{F}_{\text{thermal2}}$) independently through a sub-network. Within this sub-network, the thermal feature $\mathbf{F}_d$ and the image $\mathbf{F}_i$ are first concatenated, followed by a channel attention block to generate the weight map. The

channel attention block includes a convolutional layer, an average pooling layer, a feed-forward network layer, and a sigmoid layer to restrict the weight within the range of 0 to 1. Ultimately, the weight map is applied to augment the image feature through element-wise multiplication with the original image feature. The overall process can be formulated as:

$$\mathbf{F}_{\text{it}} = \texttt{Conv}([\mathbf{F}_{\text{i}}; \mathbf{F}_{\text{t}}]), \tag{2}$$

$$\mathbf{F}_{\text{mm}} = \sigma(\texttt{FFN}(\texttt{Avg}(\mathbf{F}_{\text{it}}))) \cdot \mathbf{F}_{\text{it}}, \tag{3}$$

where $\mathbf{F}_{\text{mm}} \in \mathbb{R}^{C \times H \times W}$ represents the fusion feature obtained by merging image and thermal inputs. By performing the aforementioned procedures, we derive $\mathbf{R}_{\text{mm}}$ and $\mathbf{I}_{\text{mm}}$. The predicted normal-light image $\mathbf{P}$ of MTRN computed as:

$$\mathbf{P}_{\text{MTRN}} = \mathbf{R}_{\text{mm}} \cdot \mathbf{I}_{\text{mm}}. \tag{4}$$

In line with the previously mentioned Retinex theory, which posits that an object's color remains constant despite variations in lighting, we not only use the Mean Squared Error (MSE) Loss to assess the difference between $\mathbf{P}_{\text{MTRN}}$ and the ground-truth normal-light image $\mathbf{P}_{\text{GT}}$, but also decompose $\mathbf{P}_{\text{GT}}$ into $\mathbf{R}_{\text{norm}}$ and $\mathbf{I}_{\text{norm}}$. This decomposition similarly guides the partitioning of the low-light image and thermal map, as illustrated in Figure 4. Particularly, we use another Retinex Encoder to encode normal-light image during the training phrase to guide the low-light image disapart. This similarity loss $\mathcal{L}_{\text{sim}}$ is the sum of the MSE loss $\mathcal{L}_{\text{R}}$ between $\mathbf{R}_{\text{mm}}$ and $\mathbf{R}_{\text{norm}}$, and the MSE loss $\mathcal{L}_{\text{I}}$ between $\mathbf{R}_{\text{mm}}$ and $\mathbf{R}_{\text{norm}}$, which can be formulated as :

$$\mathcal{L}_{\text{sim}} = \texttt{MSE}(\mathbf{R}_{\text{mm}}, \mathbf{R}_{\text{norm}}) + \texttt{MSE}(\mathbf{I}_{\text{mm}}, \mathbf{I}_{\text{norm}}). \tag{5}$$

Thermal-guided Refine Network

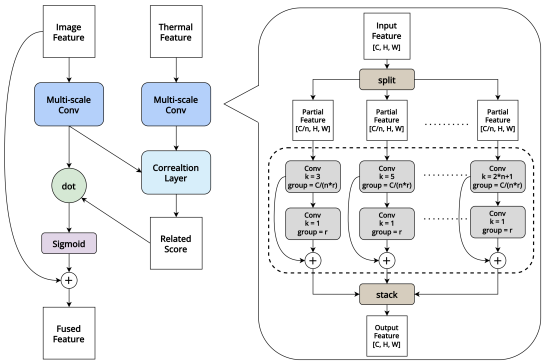

Figure 5: Details of the Multi-scale Thermal Fusion Layer.

After we have obtained the preliminary prediction result $\mathbf{P}_{\text{MTRN}}$, we first merge $\mathbf{P}_{\text{MTRN}}$ and $\mathbf{F}_{\text{SNR}}$ together, and then send it into the Thermal-guided Refine Network (TGRN) to further refine the preliminary prediction through an encoder-decoder network to obtain more accurate results. Given the complex and multi-scale characteristics of the image and thermal data $\mathbf{T} \in \mathbb{R}^{1 \times H \times W}$, it is crucial to capture features at various scales effectively prior to their fusion. Hence, as previously mentioned, in TGRN, we employ a Multi-scale Thermal Fusion Layer (MTFL) to integrate image and thermal features with diverse scales of receptive fields.

As shown in Figure 5, in MTFL, for both the image feature $\mathbf{F}_{\text{i}}$ and the thermal feature $\mathbf{F}_{\text{t}}$, we first divide the features equally into n parts along the channel dimension and use group convolutional kernels of different sizes to process the features, allowing each pixel to gather information from itself and its surrounding pixels. Then, we use another group convolution with kernel size equal to 1 to enhance those features. Finally, we concatenate all the n features along the channel dimension to integrate those information.

After obtaining the new image and thermal features $\mathbf{F}_{\text{i}'}$ and $\mathbf{F}_{d}^{'}$, to integrate thermal information into the image features, we first calculate their correlation using a correlation layer based on cross-attention (Hou et al. 2019) as follows:

$$\mathbf{T} = \texttt{softmax}\left(\frac{\mathbf{F}_{\text{i}}^{'} \cdot \mathbf{F}_{\text{t}}^{'T}}{\sqrt{d_T}}\right), \tag{6}$$

where $\mathbf{T}$ is the attention score, d_T is the dimension of $\mathbf{F}_{\text{t}}^{'}$. The score is then multiplied by $\mathbf{F}_{\text{i}}^{'}$ to integrate thermal information with the image feature, followed by a Sigmoid layer. Moreover, to prevent the gradient from vanishing, we also add a residual connection to the original image feature $\mathbf{F}_{\text{i}}$, which can be formulated as:

$$\mathbf{F}_{\text{i}} = \sigma(\mathbf{T} \cdot \mathbf{F}_{\text{i}}^{'}) + \mathbf{F}_{\text{i}}. \tag{7}$$

After feature fusion, $\mathbf{F}_{\text{i}}$ is fed into the next layer of the decoder. After repeating the above process three times, the features are sent to a prediction layer to generate the final prediction result for the normal-light image.

Loss Function

We use Mean Squared Error (MSE) as the loss function and the goal is to minimize the MSE value between the predicted RGB value to the ground truth value. The total loss $\mathcal{L}$ can be calculated as following:

$$\mathcal{L} = \mathcal{L}_{\text{MTRN}} + \mathcal{L}_{\text{TGRN}} + \mathcal{L}_{\text{SNR}} + \mathcal{L}_{\text{sim}}, \tag{8}$$

where $\mathcal{L}_{\text{MTRN}}$ is the MSE loss of MTRN, $\mathcal{L}_{\text{TGRN}}$ is the MSE loss of TGRN, $\mathcal{L}_{\text{SNR}}$ is the MSE loss of MSAM, and $\mathcal{L}_{\text{sim}}$ is the similarity loss.

Experiment

Experiment Settings

We implement ThermNet using Pytorch (Paszke et al. 2019) and train it with a 24GB RTX4090 GPU card. Throughout the training process, we resize the input images to a size of 384 $\times$ 384. For optimization, we employ Adam (Kingma and Ba 2014) with a momentum value of 0.9 and a weight decay of 1e-4. The learning rate is set to 2e-4, while the batch size and the number of training epochs are set to 8 and 100, respectively. The TLIE dataset is randomly divided into train:valid sets in an 8:2 ratio. For evaluation, we use four metrics: Peak Signal-to-Noise Ratio (PSNR) (Hore and Ziou 2010), Structural Similarity Index (SSIM) (Wang et al. 2004), Multi-Scale Structural Similarity Index (MSSM) (Wang et al. 2003) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018). We use a pre-trained AlexNet (Krizhevsky, Sutskever, and Hinton 2012) as the feature extraction backbone in LPIPS.

Method	PSNR↑	SSIM↑	MSSM↑	LPIPS↓	Param
Image-only Methods					
EFINet	17.67	0.420	0.645	0.684	**<u>0.13M</u>**
ChebyLighter	17.56	0.413	0.632	0.619	**0.18M**
DeepLPF	17.63	0.425	0.646	0.671	1.8M
DNF	18.95	0.441	0.698	0.731	2.8M
MIRNet	19.16	0.440	0.706	0.702	31.8M
SNR-Net	19.22	0.451	0.706	0.610	39.1M
LEDNet	19.33	0.440	0.707	0.742	7.4M
Restormer	19.41	0.450	0.708	0.635	26.1M
Retinexformer	19.42	0.446	0.714	**0.605**	1.6M
Multimodal Methods					
eSL-Net	18.39	0.427	0.659	0.692	0.19M
ELIE	**19.50**	**0.452**	**0.717**	0.658	220M
ThermNet (ours)	**<u>19.81</u>**	**<u>0.468</u>**	**<u>0.730</u>**	**<u>0.556</u>**	0.51M

Table 2: Comparison of ThermNet with different SOTA methods from LIE task on TLIE dataset. The top and second best results are highlighted in **<u>underline</u>** and **bold**.

Method	PSNR↑	SSIM↑	MSSM↑	LPIPS↓
EFINet	17.70 (<u>0.03</u>)	0.423 (<u>0.003</u>)	0.653 (<u>0.008</u>)	0.706 (**0.022**)
DeepLPF	19.74 (<u>1.68</u>)	0.457 (<u>0.043</u>)	0.793 (<u>0.044</u>)	0.565 (<u>0.133</u>)
DNF	19.29 (<u>0.34</u>)	0.443 (<u>0.002</u>)	0.708 (<u>0.010</u>)	0.690 (<u>0.041</u>)
SNR-Net	19.27 (<u>0.05</u>)	0.453 (<u>0.002</u>)	0.704 (**0.002**)	0.590 (<u>0.020</u>)
LEDNet	19.55 (<u>0.22</u>)	0.447 (<u>0.007</u>)	0.721 (<u>0.014</u>)	0.681 (<u>0.061</u>)

Table 3: Experimental results of previous SOTA methods enhanced with our MTFL module. Performance improvements are shown in parentheses, where <u>underline</u> numbers represent increased performance and **bold** represent degraded.

Quantitative Anaysis

We first conduct an experiment that compares our ThermNet with previous SOTA methods on our TLIE dataset. Here, we choose LIE methods over the past four years, including MIRNet [ECCV2020] (Zamir et al. 2020), DeepLPF [CVPR2020] (Moran et al. 2020), SNR-Net [CVPR2022] (Xu et al. 2022), Restormer [CVPR2022] (Zamir et al. 2022), EFINet [TCSVT2022] (Liu, Wu, and Wang 2022), ChebyLighter [MM2022] (Pan et al. 2022), LEDNet [ECCV2022] (Zhou, Li, and Change Loy 2022), DNF [CVPR2023] (Jin et al. 2023), Retinexformer [ICCV2023] (Cai et al. 2023), we also include two multimodal event-based LIE methods eSL-Net [ECCV2020] (Wang et al. 2020) and ELIE [TMM2023] (Jiang et al. 2023), where we replace the input event with our thermal map. All experimental results are shown in Table 2. ThermNet achieves SOTA performance on all metrics in the TLIE. Compared to the unimodal LIE method Retinexformer, ThermNet can outperform it by a significant margin especially in PSNR (19.81 vs. 19.42), SSIM (0.468 vs. 0.446), and LPIPS (0.556 vs. 0.605). Compared to ELIE, ThermNet not only has higher accuracy, but our parameters are also fewer than it.

In Figure 1, we draw the relationship between FLOPs, parameters, and PSNR of all methods, and it can be found that our method has fewer parameters and fewer FLOPs compared to other methods, but still achieving the highest accuracy, proving the efficiency and effectiveness of our method. In Figure 6, we show the PSNR score of different algorithms in the TLIE validation set at different epochs. It can be observed that our method converges at a relatively slow rate, but as training continues, it gradually shows its advantage in accuracy, ultimately outperforming the other methods. This highlights the effectiveness of our algorithm.

To further reveal the necessity of incorporating thermal information into the LIE task, we try to enhance the previous SOTA methods by integrating thermal information fusion into their models. The goal is to observe whether this integration could further improve accuracy compared to using only RGB images. We test several methods from Table 2 and enhance them with our MTFL module to incorporate thermal information into them. The results are shown in Table 3. We observe that in the vast majority of cases, MTFL brings substantial performance improvements to different methods. This not only validates the effectiveness of the MTFL module we designed but also shows the importance of introducing thermal information for low-light image enhancement.

MTRN	MSAM	MTFL	PSNR↑	SSIM↑	MSSM↑	LPIPS↓
✗	✗	✗	18.53	0.439	0.695	0.652
✓	✗	✗	19.46	0.455	0.721	0.595
✗	✓	✗	19.26	0.447	0.701	0.602
✗	✗	✓	19.48	0.450	0.714	0.589
✓	✓	✗	19.63	0.457	0.724	0.572
✓	✗	✓	19.74	0.458	0.726	0.571
✗	✓	✓	19.55	0.451	0.713	0.568
✓	✓	✓	**19.81**	**0.468**	**0.730**	**0.556**

Table 4: Ablation studies on different modules in ThermNet.

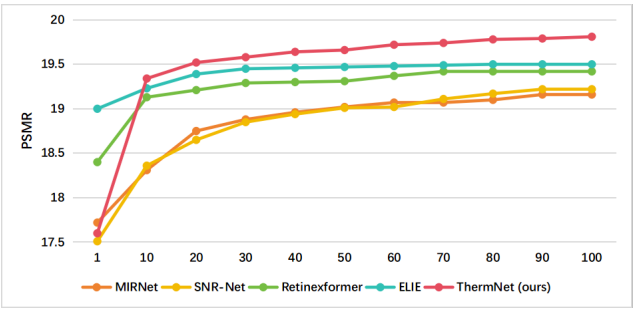

Figure 6: Trending of PSNR on validation set.

Ablation Study

We assess the effect of each module on ThermNet by disabling one or multiple components and evaluating performance on the TLIE test set, as displayed in Table 4. First, it can be observed that by reconstructing the SNR map and fusing it with the RGB image, we are able to enhance the model's performance, which reveals not only the importance

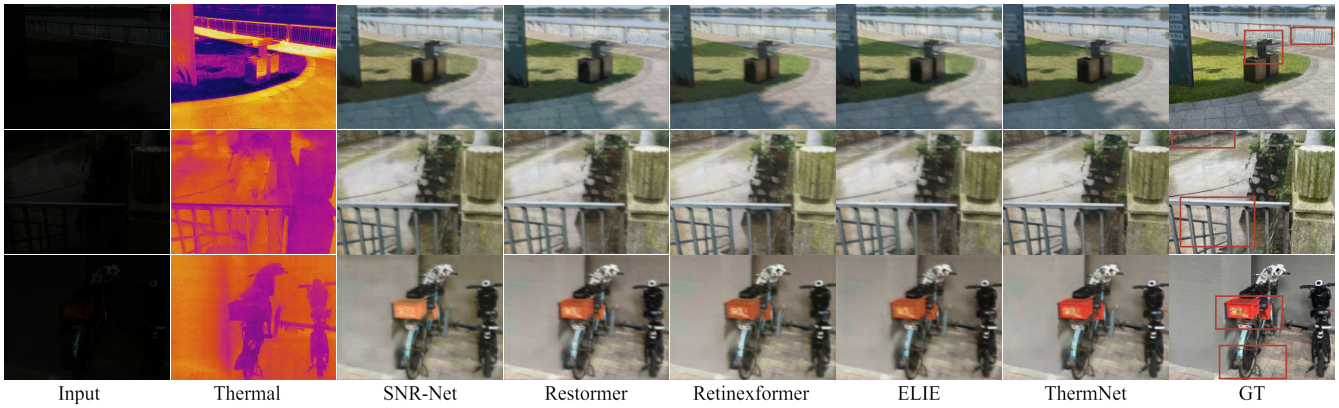

Figure 7: Results on TLIE dataset. Compared with best-performing baselines, our method ThermNet are closest to the ground truth (GT), which effectively improves image visibility while preserving correct color, especially in extremely dark areas.

of thermal information in the LIE task but also the effectiveness of our MSAM. Furthermore, when using only one module, MTFL has the most significant impact among the three modules. Compared to the basic network without any specific module, it increases the PSNR accuracy by 0.95. This demonstrates the effectiveness of the feature fusion approach employed by MTFL, and meanwhile proving that introducing thermal information into the LIE task is useful. MTRN is the second most useful module, while MSAM, although not as significant as the other two, still contributes to improving the model accuracy to some extent. The configuration that integrates all three modules achieves the best accuracy, demonstrating that every component in ThermNet contributes positively to overall performance.

Qualitative Analysis

To more intuitively compare the performance differences between our approach and the previous methods, we select several samples from the test set and visualize the reconstruction results of each model in Figure 7. We highlight some significant differences from the red boxes. It can be observed that, compared to previous methods, our approach performs better on the edges of objects, and our model's reconstruction results are closest to the ground truth, especially in particularly dark areas. For instance, in the first sample, our method restores the outline of the trash can more clearly, especially its upper part, while other methods such as ELIE and Retinexformer are comparatively blurrier. Additionally, for the railing behind the trash can, with the help of a clear thermal map, the railing restored by our method is more distinct and identifiable. In the second sample, compared to other methods, our method restores the railing and the details of the pile of rocks in the upper left corner of the image much more closely to the original image. Similarly, in the third sample, the clarity of the rear wheel outline of the bicycle is much higher in our method compared to others. While in the color restoration of the red basket on the bike, our method's color is closest to the original color (red), while other methods restored the color as more yellowish.

To further compare the performance of the models, we take a set of photos in real low-light environments, both indoors and outdoors, for the models to restore, which is shown in Figure 8. We can observe that compared to unimodal methods, our multimodal approach has a significant advantage in terms of clarity. All of the above examples demonstrate that by leveraging thermal information, our method exhibits better robustness for low-light images under different lighting and environmental conditions.

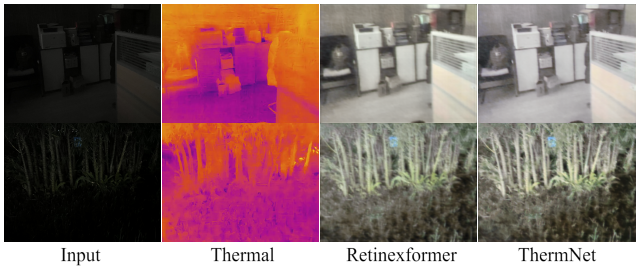

Figure 8: Comparison of reconstruction results of several best-performing baselines in real low-light environments.

Conclusion

This study focuses on multimodal low-light image enhancement (MLIE) with thermal maps. To support ongoing research in this domain, we develop a novel and extensive dataset named TLIE, comprising 1,113 samples, each containing a low-light image, a standard-light image, and an associated thermal map. Furthermore, we introduce an innovative network named ThermNet that integrates feature fusion of RGB images with thermal maps to enhance low-light image enhancement. Extensive experiments and thorough analyses show that ThermNet surpasses current single-model based LIE methods when incorporating thermal data from the TLIE dataset, thus validating the importance of using thermal maps in LIE. Additionally, by integrating our custom-designed module with existing methods, we demonstrate the efficacy and versatility of our approach, highlighting its potential for seamless integration with other models.

Acknowledgments

This work is supported by Guangxi Science and Technology Major Project AA23062030.

References

Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; and Zhang, Y. 2023. Retinexformer: One-stage Retinex-based Transformer for Low-light Image Enhancement. *arXiv preprint arXiv:2303.06705*.

Cao, Y.; Tong, X.; Wang, F.; Yang, J.; Cao, Y.; Strat, S. T.; and Tisse, C.-L. 2023. A deep thermal-guided approach for effective low-light visible image enhancement. *Neurocomputing*, 522: 129–141.

Chen, C.; Chen, Q.; Do, M. N.; and Koltun, V. 2019. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3185–3194.

Chen, C.; Chen, Q.; Xu, J.; and Koltun, V. 2018. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and pattern recognition*, 3291–3300.

Choi, Y.; Kim, N.; Hwang, S.; Park, K.; Yoon, J. S.; An, K.; and Kweon, I. S. 2018. KAIST multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3): 934–948.

FLIR, T. 2022. FREE Teledyne FLIR Thermal Dataset for Algorithm Training. https://www.flir.com/oem/adas/adas-dataset-form/.

Guo, X.; Li, Y.; and Ling, H. 2016. LIME: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2): 982–993.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *International Conference on Pattern Recognition*, 2366–2369. IEEE.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2019. Cross attention network for few-shot classification. *Advances in neural information processing systems*, 32.

Jiang, Y.; Wang, Y.; Li, S.; Zhang, Y.; Zhao, M.; and Gao, Y. 2023. Event-based low-illumination image enhancement. *IEEE Transactions on Multimedia*.

Jin, X.; Han, L.-H.; Li, Z.; Guo, C.-L.; Chai, Z.; and Li, C. 2023. DNF: Decouple and Feedback Network for Seeing in the Dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18135–18144.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

Land, E. H. 1977. The retinex theory of color vision. *Scientific American*, 237(6): 108–129.

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

Li, D.; Wang, Z.; Chen, Y.; Jiang, R.; Ding, W.; and Okumura, M. 2024. A Survey on Deep Active Learning: Recent Advances and New Frontiers. *IEEE Transactions on Neural Networks and Learning Systems*.

Liang, G.; Chen, K.; Li, H.; Lu, Y.; and Wang, L. 2024. Towards Robust Event-guided Low-Light Image Enhancement: A Large-Scale Real-World Event-Image Dataset and Novel Approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23–33.

Liu, C.; Wu, F.; and Wang, X. 2022. Efinet: Restoration for low-light images via enhancement-fusion iterative network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8486–8499.

Liu, Q.; Li, X.; He, Z.; Li, C.; Li, J.; Zhou, Z.; Yuan, D.; Li, J.; Yang, K.; Fan, N.; et al. 2020. LSOTB-TIR: A large-scale high-diversity thermal infrared object tracking benchmark. In *Proceedings of the 28th ACM international conference on multimedia*, 3847–3856.

Moran, S.; Marza, P.; McDonagh, S.; Parisot, S.; and Slabaugh, G. 2020. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12826–12835.

Pan, J.; Zhai, D.; Bai, Y.; Jiang, J.; Zhao, D.; and Liu, X. 2022. ChebyLighter: Optimal Curve Estimation for Low-light Image Enhancement. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1358–1366.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, 746–760. Springer.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 567–576.

Wang, B.; He, J.; Yu, L.; Xia, G.-S.; and Yang, W. 2020. Event enhanced high-quality image recovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 155–171. Springer.

Wang, R.; Xu, X.; Fu, C.-W.; Lu, J.; Yu, B.; and Jia, J. 2021. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9700–9709.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wang, Z.; Li, D.; and Jiang, R. 2024. Diffusion Models in 3D Vision: A Survey. *arXiv preprint arXiv:2410.04738*.

Wang, Z.; Li, D.; Li, G.; Zhang, Z.; and Jiang, R. 2024. Multimodal low-light image enhancement with depth information. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4976–4985.

Wang, Z.; Simoncelli, E. P.; Bovik, A. C.; and et al. 2003. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, 1398–1402. Ieee.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*.

Xu, X.; Wang, R.; Fu, C.-W.; and Jia, J. 2022. SNR-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17714–17724.

Yang, W.; Wang, W.; Huang, H.; Wang, S.; and Liu, J. 2021. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30: 2072–2086.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 5728–5739.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2020. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 492–511. Springer.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, Y.; Guo, X.; Ma, J.; Liu, W.; and Zhang, J. 2021. Beyond brightening low-light images. *International Journal of Computer Vision*, 129: 1013–1037.

Zhang, Y.; Zhang, J.; and Guo, X. 2019. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1632–1640.

Zhou, S.; Li, C.; and Change Loy, C. 2022. Lednet: Joint low-light enhancement and deblurring in the dark. In *European Conference on Computer Vision*, 573–589. Springer.