Learning to Restore Low-Light Images via Decomposition-and-Enhancement

Ke Xu, Xin Yang, Baocai Yin, Rynson W.H. Lau

1. Introduction

Low-light imaging is very popular, for various purposes, e.g., night-time surveillance and personal scenery imaging at sunset. However, the visibility of low-light images in the standard RGB (sRGB, 24 bits/pixel) space does not match with human perception, due to quantization. This low visibility hinders vision tasks (e.g., object detection [31] and tracking [8]), or image editing tasks (e.g., image matting [45]). Hence, recovering low-light images is essential.

Typical image enhancement methods [46, 51, 24, 7, 40, 34, 48, 4] propose to recover low-light images to match with human perception. These methods rely on users to have good photographic skills in taking images with low noise, so that these methods can focus on learning to manipulate the tones, colors or contrasts of the images. As such, they cannot be used to enhance majority of the practical low-light images with noise, which are taken by casual users. Figure 1 shows one example, where image contents are not only buried by low pixel intensity values, but also disrupted by noise, due to the inherent low signal-to-noise ratio (SNR) at low light [6]. Existing enhancement methods may either enhance both the noise and scene details (Figure 1(b, f)), or fail to recover the low visibility of low-light images (Figure 1(e)). In addition, these enhanced images still have low SNRs, providing limited useful contextual information for detecting noise from scene details. Hence, they fail existing image denoising methods [11, 49, 50, 27, 37, 32, 19].

In this paper, we address the low-light sRGB image enhancement problem, which involves two issues: image enhancement as well as denoising. Our motivation is based on two observations. First, the image low-frequency layer preserves more information, e.g., objects and colors, and is less affected by noise (Figure 1(e)) than the image high-frequency layer (Figure 1(d)). This suggests that it is easier to enhance the low-frequency image layer than to directly enhance the whole image. Second, the very low intrinsic dimensionality of image primitives makes it possible for neural networks to learn a full knowledge of image primitives [29, 41]. Hence, given the low-frequency informa-
tion of primitives, it is possible for a network to reconstruct the whole primitives by inferring the corresponding high-frequency information. With such a prior, we can then learn to enhance high-frequency details from the recovered low-frequency layer.

These two insights inspire us to learn a frequency-based low-light image decomposition-and-enhancement model. To this end, we propose a novel neural network that leverages an Attention to Context Encoding (ACE) module to adaptively select low-frequency information for recovering the low-frequency layer and noise removal in the first stage, and select high-frequency information for detail enhancement in the second stage. We also propose a Cross Domain Transformation (CDT) module to leverage multi-scale frequency-based features for noise suppression and detail enhancement in the two stages. As shown in Figure 2, our method can enhance the noisy low-light sRGB image with contents/details recovered and noise suppressed.

In summary, the main contributions of this work are:

1. We propose a novel frequency-based decomposition-and-enhancement model for enhancing low-light images. It first recovers image contents in the low-frequency layer while suppressing noise, and then recovers high-frequency image details.

2. We propose a network, with an Attention to Context Encoding (ACE) module to decompose the input image for adaptively enhancing the high-/low-frequency layers and a Cross Domain Transformation (CDT) module for noise suppression and detail enhancement.

3. We prepare a low-light image dataset with real noise and corresponding ground truth images, to facilitate the learning process. Extensive experiments verify the superior performance of the proposed method over the state-of-the-art approaches.

2. Related work

Low-light image enhancement. A line of methods enhance low-light images using different image-to-image regression functions. Represented by histogram equalization [36] and gamma correction, global and local contrast enhancement operators are proposed based on detecting semantic regions (e.g., face and sky) [25], matching region templates [23] or contrast statistics in image boundaries and textured regions [38]. Advanced deep learning based methods learn the mapping functions from high-quality user retouched images or images taken using high-end cameras, using bilateral learning [15], intermediate HDR supervision [46], adversarial learning [24, 7], or reinforcement learning [34, 48]. Another line of works are retinex-based image enhancement methods [20, 14, 51, 5, 40, 47], which decompose the input low-light image into illumination and reflectance, and then enhance the illumination of the image.

However, existing enhancement methods may fail to recover low-light images, due to their low SNRs, as shown in Figure 2. The key reason is that these methods [24, 34, 7, 48, 46] typically assume the images to be taken by photographical experts with insignificant noise levels. Hence, they are unable to enhance noisy low-light images.

Recently, there are also some enhancement methods [6, 22] proposed to directly retouch the camera raw data into high quality output images. Particularly, Chen et al. [6] proposed to learn raw-to-image models to generate noise-suppressed, enhanced images from noisy raw images. However, models trained on the raw domain cannot be applied to regular sRGB images, which is the most widely adopted color space [10], as the linear raw data is significantly different from the non-linear RGB data [44]. Besides, raw data is usually unavailable due to the lack of expertise or unknown protocols. In this paper, we focus on enhancing noisy low-light sRGB images.

Image denoising. Single image denoising is an active research topic in computer vision, and it often functions as pre-/post-processing for other vision tasks. Many methods have been developed based on image priors such as self-similarity [3, 11], sparsity [13, 30], and low rank [18, 43]. Deep learning has also been widely applied to the denoising problem [33, 49, 50, 27, 37, 32]. These denoisers typically learned from synthetic datasets that assumed additive, white or Gaussian noise. They often fail to remove real noise, which exhibits different patterns. Recent works attempted to improve the performances of denoisers in denoising real
images, by synthesizing noise in the raw data domain [2], constructing real image dataset [1], developing joint training strategy of both synthetic and real images [19], or unsupervised learning [28].

However, it is non-trivial to remove noise from low-light images simply by pre-/post-processing with existing denoising methods. On the one hand, low pixel values make it difficult to provide sufficient contextual information for detecting/removing noise before enhancing the low-light images. On the other hand, noise can be unpredictably amplified after applying existing enhancement methods, producing images that still have low SNRs and hence difficult for further denoising. To address this limitation, we propose in this paper to learn a deep enhancement model to enhance the low-light images while removing noise, in an end-to-end recurrent manner.

3. Proposed Model

Our method is inspired by two observations. First, it is easier to enhance the low-frequency layer of a noisy low-light image, compared to directly enhancing the whole image. This is because noise in the low-frequency layer is easier to detect and then suppress. Image illumination/colors can then be properly estimated by analyzing the global properties of the image low-frequency layer. Second, it is known that primitive parts of natural images, e.g., edges and corners, have very low intrinsic dimensionality [29]. Such low dimensionality implies that a small number of image examples are sufficient to represent the image primitives well [41]. Hence, given the low-frequency information of the primitives, we may be able to infer the corresponding high-frequency information.

Based on these two observations, our proposed model, as shown in Figure 3, has two main stages. In the first stage, we propose to learn a low-frequency image enhancement function $C(\cdot)$, and then an amplification function $A(\cdot)$ for color recovery. By jointly modeling the mapping from $C(\cdot)$ to $A(\cdot)$, the network does not have to learn both global information (e.g., illumination) and local information (e.g., color) at the same time, resulting in a more effective enhancement. Formally, given a low-light sRGB image $I$, the first stage enhancement can be written as:

$$I^a = \alpha A(C(I)) \cdot C(I),$$

where $I^a$ is the amplified low-frequency layer. Note that $A$ is different from the illumination map in retinex-based methods, as we estimate a relative amplification map to a learnable global ratio $\alpha$ from the enhanced content $C$. In other words, $\alpha A(\cdot)$ can be interpreted as an error map that enhances $C$ in the self-attention manner.

In the second stage, we propose to learn high-frequency detail enhancement function $D(\cdot)$, based on $I^a$ from the first stage, instead of directly restoring the high-frequency details from the original input image $I$, which is noisy. $D(\cdot)$ is then modeled in a residual manner, and the final enhanced image can be obtained as:

$$I^c = I^a + D(I^a).$$

Our model uses two novel modules, the Attention to Context Encoding (ACE) module and the Cross Domain Transformation (CDT) module. They are explained below.

3.1. ACE Module

The goal of the ACE module is to learn frequency-aware features for image decomposition. To do this, we extend the non-local operation [42], originally proposed for encoding long-range relations, to select frequency adaptive contextual information. Figure 5 shows the block diagram.

We use the first ACE module in Figure 3 for explanation. Given the input features $x_{in} \in \mathbb{R}^{H \times W \times C}$, we first use two
groups of dilated convolutions (with kernel size/dilation rate of 1/1 and 3/2), denoted as \( f_{d1} \) and \( f_{d2} \), to extract features in different receptive fields. We then compute a contrast-aware attention map \( C_a \) between these two features as:

\[
C_a = \text{sigmoid}(f_{d1}(x_{in}) - f_{d2}(x_{in})).
\]

\( C_a \) indicates the pixel-wise relative contrast information, where pixels of high contrasts are regarded as belonging to the high-frequency layer. We then compute the inverse map \( \overline{C}_a = 1 - C_a \) to select features from \( x_{in} \) to represent the low-frequency contents as: \( x_c = \overline{C}_a \cdot x_{in} \). We further shrink the selected features \( x_c \) via max-pooling to obtain compact features \( x_c^1 \) and to reduce GPU memory and computations for establishing the non-local pixel-to-pixel dependence. Formally, given \( x_c^1 \in R^{H' \times W' \times C} \), the non-local context encoding process can be written as:

\[
x_c^2 = g(x_c^1)^\top \times h(x_c^2) \times f(x_c^1)^\top,
\]

where \( g, h, f \) represent groups of operations (convolution, reshaping and matrix transpose) that first compute a pixel affinity table \( M \in R^{H' \times W' \times H' \times W'} \) and then compute non-locally enhanced features \( x_c^2 \) by considering the relations of each pixel to all other pixels. Finally, we obtain the frequency-aware non-locally enhanced features \( x_{out} = \text{Unpool}(x_c^2) + x_c \) in a residual manner to facilitate the learning process. Note that the two ACE modules in Figure 3 share their weights. The second ACE module uses the contrast-aware attention map \( C_a \), instead of the inverse map \( \overline{C}_a \), to learn the image details from the features representing the high-frequency layer. Figure 6 shows two ACE attention maps (\( C_a \) from the first stage and \( C_a \) from the second stage) and their corresponding decomposed feature maps (\( x_c^1 \) from the first stage and \( x_c^2 \) from the second stage).

### 3.2. CDT Module

A good understanding of the global properties of low-light images can help recover the lighting and image contents. To do this, we propose the CDT module, as shown in Figure 7, to increase the receptive fields while bridging the gap between features in the low-light domain and in the enhanced domain. Sharing a similar spirit as [39] in increasing the receptive fields for more global information, the CDT module is specially designed to concurrently address the domain gap problem, i.e., frequency-aware features extracted in the noisy low-light domain versus those in the enhanced domain.

Specifically, in the first stage, the noisy features from the encoder \( x_{en} \) are first spatially reweighed via the self-derived inverse contrast-aware map \( \overline{C}_a \) to filter out high contrast information, before concatenating with features \( x_{de} \) from the corresponding decoder. We then compute global scaling vectors \( v \) from the concatenated features \( [x_{en}, x_{de}] \), for adaptively re-scaling the features from different domains in a channel-wise manner. In the second stage, we use the contrast-aware attention map \( C_a \), instead of the inverse map \( \overline{C}_a \), to learn image details, similar to the ACE module.

### 3.3. Proposed Dataset

To facilitate the learning of the proposed model, we have prepared a new low-light dataset of real noisy low-light and ground truth sRGB image pairs.

**Noise in low-light.** We prepare our training data based on the SID dataset [6], which consists of raw data and ground truth image pairs. This raw data was collected...
when imaging in low-light with short exposure time (typically 0.1s or 0.04s). Their corresponding ground truth images were taken with long exposure time (typically 10s or 30s), where noise is negligible. However, the linear camera raw data is significantly different from the non-linear sRGB data, particularly in terms of noise [2] and image intensity [46]. As a result, models trained on raw data cannot be directly applied to sRGB images. To address this problem, we have considered several key steps (i.e., exposure compensation, white balance and de-linearization) in the image formation pipeline, and manipulated their operations in order to model real-world noisy low-light sRGB images taken from different cameras.

**Exposure compensation.** Auto-exposure algorithms aim to automatically determine the exposure time and camera gain based on the light intensity perceived by the sensor. They are usually black-boxes and vary across cameras. To augment the diversity of this exposure time, we randomly sample the exposure compensation value from the range of $[0EV, 2EV]$ at intervals of $0.5EV$.

**White balance.** White balance algorithms aim to correct unrealistic casts via estimating the per-channel gain [16]. They are also unknown and vary across cameras. We augment it by randomly choosing the color temperature from the range of $[2100K, 4000K]$, which represents the color temperatures of typical household lighting and Sunrise/Sunset lighting, according to the Kelvin temperature color chart [9].

**De-linearization.** As the non-linearity introduced by the camera response function varies across cameras and is difficult to reverse-engineer [17], we apply the gamma function as the de-linearization function, as suggested in [12].

Using the above settings, we have produced a total of 4,198 image pairs for training and 1,196 image pairs for testing. Experimental results in Figures 9 and 10 show that the proposed network trained on our data can generalize well on images from other image formation pipelines.

**3.4. Training**

**Loss function.** We use L2 loss to measure the reconstruction accuracy in the two-stage training process. Specifically, in the first stage, to encourage our network to focus on predicting the low frequency components of the input image, we prepare the corresponding ground truth, denoted as $I^g_t$, by using the guided filter [21] to filter out the high-frequency details while maintaining the main structures and contents of the ground truth image. Formally, the reconstruction loss can be written as:

$$L_{acc} = \lambda_1 \left \| C - I^g_t \right \|_2 + \lambda_2 \left \| I^c - I^g_t \right \|_2,$$  \hspace{1cm} (5)

where $C$, $I^c$, $I^g_t$, $I^g$ are the reconstructed image content, the recovered image, ground truth of the low-frequency layer, and ground truth of the enhanced image, respectively. $\lambda_1$ and $\lambda_2$ are balancing parameters.

We also incorporate the perceptual loss by comparing the VGG feature distances of $I^c$ and $I^g$, using L1 loss, as:

$$L_{vgg} = \lambda_3 \left \| \Phi (I^c) - \Phi (I^g) \right \|_1,$$  \hspace{1cm} (6)

where $\Phi$ is the VGG net, and $\lambda_3$ is a balancing parameter.

**4. Experiments**

We have implemented the proposed model in the PyTorch framework [35], and tested it on a PC with an i7 4GHz CPU and a GTX 1080Ti GPU. As we train our model from scratch, the network parameters are initialized randomly, except the learnable amplification ratio $\alpha$, which is initialized to 1. Standard augmentation strategies, i.e., scaling, cropping, and horizontal flipping, are adopted. During training, we randomly crop patches of resolutions $512 \times 384$ from the scaled images of resolution $2048 \times 1536$. For loss minimization, we adopt the ADAM optimizer [26] for 400 epochs, with an initial learning rate of $3e^{-4}$ and divided by 10 at the 250th epoch. $\lambda_1$, $\lambda_2$ and $\lambda_3$ are set to 1, 1, and 0.1, respectively. It takes 0.33s for the proposed network to process one image of resolutions $1024 \times 768$.

To evaluate the performance of the proposed method on enhancing low-light images, we quantitatively and visually compare our method to 9 state-of-the-art enhancement methods with available codes, including JieP [5], LIME [20], WVM [14], DSLR [24], CAPE [25], DRHT [46], DeepUPE [40], HDRCNN [12] and SID [6]. We use PSNR and SSIM for quantitative measurement.

**4.1. Comparing to State-of-the-Arts**

**Visual comparisons.** We first visually compare results of the proposed method to the state-of-the-art enhancement methods. Figure 8 shows the results of different methods on three input low-light images (a, m, A), which were taken by a Sony camera. We can see that WVM [14] and DeepUPE [40] fail to enhance these images (c, d, o, p, C, D). Since they are based on decomposing the input image into reflectance and illumination, when an input image is of low-light, they are unable to decompose it accurately. LIME [20] can enhance the images (f, r, F), as it directly estimates the illumination map without decomposing the input image. However, it enhances both details and noise together. Similarly, the gamma correction based method CAPE [25] also jointly enhances the details and noise together (e, q, E). DRHT [46] fails to enhance the noisy low-light images (h, t, H), as noise can deteriorate both the HDR reconstruction and tone mapping processes. DSLR [24] is trained to regress a low-quality image into a high-quality one. While it can somewhat enhance the images, it fails to remove noise (i, u, I). Since the original SID [6] model
Figure 8. Visual results of state-of-the-art methods and ours on input low-light images (a, m, A). Red boxes indicate the noisy regions where most existing methods fail. The input images were taken by a Sony camera.

(trained on raw domain) cannot be directly applied to sRGB images, we re-train it on the sRGB images. We can see that the SID model tends to remove noise and details, resulting in blurred images (j, v, J). In contrast, our results (l, x, L) show that the proposed method can successfully enhance the image content and details while suppressing noise.

Figure 9 shows results of another three input low-light images (taken by an iPhone camera). While state-of-the-art methods generally fail to remove noise and enhancing contents/details at the same time, our method produces visually more convincing results, even for the more challenging textured images (l, x). Figures 8 and 9 demonstrate the good generalization ability of the proposed model/dataset on images taken by different types of cameras.

Quantitative comparisons. We have also quantitatively compared our method to the state-of-the-art enhancement methods. As shown in Table 1, the proposed method outperforms these existing enhancement methods by a large margin. Note that we have also pre-processed the input images before feeding them to two methods [14, 5], by amplifying these image pixel intensities with pre-defined ratios as in [6] or by applying histogram equalization. However, the results are the same as those without pre-processing. This indicates that enhancing noisy low-light images via decomposing images into reflectance and illumination is not suitable. In contrast, our frequency-based decomposition-and-enhancement can successfully decouple the image enhancement and denoising problem.

We also compare our method with SID [6], which was originally proposed to enhance low-light images in the raw domain, in both sRGB and raw domains. Specifically, in the sRGB domain, we apply two strategies: directly using the original SID model trained on raw images (denoted as SID), and using a retrained SID model on sRGB images in our training set (denoted as SID*). In the raw domain, we retrain our model using the raw data. We can see that our method outperforms SID [6] in both sRGB and raw domains. We further compare our method to the newest method [40] in both sRGB (retrained on our dataset) and raw domains. These results show that our model is more effective in enhancing low-light images with noise, than directly learning the image-to-image [6] or image-to-illumination [40] regression models.

Finally, we compare our method to different combinations of existing enhancement and denoising methods. Specifically, we choose one classic denoising method B-
Figure 9. Visual results of state-of-the-art methods and ours on input low-light images (a, m, A). Red boxes indicate the noisy regions where most existing methods fail. The input images were taken by an iPhone camera. Results of our method in here as well as in Figure 8 demonstrate the generalization ability of the method on different camera types.

M3D [11] and one recent deep learning based denoising method xDnCNN [27] to pre-/post-process the low-light images (in the test set) before/after they are processed by enhancement method LIME [20]. We choose LIME [20] as it has the third best performance among the existing methods in Table 1. Although SID* [6] and DeepUPE* [40] have better performance, they are already trained on our dataset to remove noise. Hence, we do not use them here. Table 2 shows the results. We can see that directly applying existing denoising methods as a pre-/post-processing step to enhancement methods does not work well. As noise is already deeply buried into the image contents and details in low-light images, separately enhancing and denoising these images do not perform well. Instead, we suppress the noise in the low-frequency layer and then enhance the contents and details adaptively, producing better performances. Figure 10 shows some visual examples of combining existing enhancement and denoising methods. We can see that denoising followed by enhancement produces blurry results (e, f), due to the significant removal of image details in the denoising step. Although enhancement followed by denoising can produce relatively sharper results (g, h) in comparison to (e, f), respectively, the results are more noisy as both noise and details are enhanced in the enhancement step. It is also interesting to note that none of these methods can recover the colors (caused by noise) well, e.g., the purplish color of the tree. In contrast, our method can produce a sharp image (d), with noise suppressed and color recovered.
As a future work, we are interested in extending the proposed method against state-of-the-art methods. We also train our model by directly using the output of the first stage (denoted as $I_f^* \rightarrow I^g$), instead of using the ground truth of the low-frequency layer. Results are shown in the 6th and 7th rows. It shows the advantage of learning a two-stage model over Single Shot. We can also see that using ground truth of the low-frequency layer to supervise the first stage produces better results than using the ground truth images, which verifies the importance of learning the decomposition-and-enhancement model.

### 5. Conclusion and Future Work

In this paper, we have studied the noisy low-light image enhancement problem. We have observed that noise affects images differently in different frequency layers. Based on this observation, we propose a novel frequency-based image decomposition-and-enhancement model to adaptively enhance the image contents and details in different frequency layers, while at the same time suppressing noise. We have also presented a network with the proposed Attention to Context Encoding (ACE) module for adaptively enhancing the high and low frequency layers, and Cross Domain Transformation (CDT) module for noise suppression and detail enhancement. To train our model, we have prepared a new low-light image dataset. Finally, we have conducted extensive experiments to verify the effectiveness of our method against state-of-the-art methods.

Our method does have limitations. It may fail in scenes with small objects, in which our network may not be able to extract meaningful contextual information from the surrounding areas in order to recover the contents, as shown in Figure 11. As a future work, we are interested in extending our enhancement model to consider semantic layouts of the scenes and using generative adversarial learning for synthesizing image details.

**Acknowledgement.** This work was partly supported by NNSFC Grants 91748104, 61972067, 61632006, U1811463, U1908214, 61751203; and the National Key Research and Development Program of China, Grant 2018AAA0102003.
References


