ENHANCEMENT OF DARK IMAGES AND VIDEOS VIA SET-BASED LEARNING

KARANLIK GÖRÜNTÜ VE VİDEOLARIN KÜME TABANLI ÖĞRENME YÖNTEMİYLE İYİLEŞTİRİLMESİ

AHMET SERDAR KARADENİZ

ASSOC. PROF. DR. MEHMET ERKUT ERDEM

Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment to the Requirements

for the Award of the Degree of Master of Science

in Computer Engineering

September 2020

ABSTRACT

ENHANCEMENT OF DARK IMAGES AND VIDEOS VIA SET-BASED LEARNING

Ahmet Serdar Karadeniz

Master of Science, Computer Engineering Department

Supervisor: Assoc. Prof. Dr. Mehmet Erkut ERDEM

September 2020, 71 pages

Capturing images under extremely low-light conditions poses significant challenges for the standard camera pipeline. Images become too dark and too noisy, which makes traditional enhancement techniques almost impossible to apply. Recently, learning-based approaches have shown very promising results for this task since they have substantially more expressive capabilities to allow for improved quality. Motivated by these studies, in this thesis, we aim to leverage burst photography to boost the performance and obtain much sharper and more accurate RGB images from extremely dark raw images. The backbone of our proposed framework is a novel coarse-to-fine network architecture that generates high-quality outputs progressively. The coarse network predicts a low-resolution, denoised raw image, which is then fed to the fine network to recover fine-scale details and realistic textures. To further reduce the noise level and improve the color accuracy, we extend this network to a permutation invariant structure so that it takes a burst of low-light images as input and merges information from multiple images at the feature-level. Our experiments demonstrate that our approach leads to perceptually more pleasing results than the state-of-the-art methods by producing more detailed and considerably higher quality images.

Keywords: Computational Photography, Low-light Imaging, Image Denoising, Burst Images

ÖZET

KARANLIK GÖRÜNTÜ VE VİDEOLARIN KÜME TABANLI ÖĞRENME YÖNTEMİYLE İYİLEŞTİRİLMESİ

Ahmet Serdar KARADENİZ

Yüksek Lisans, Bilgisayar Mühendisliği

Danışman: Doç. Dr. Mehmet Erkut ERDEM

Eylül 2020, 71 sayfa

Aşırı düşük ışık koşullarında görüntü yakalamak, standart kamera hattı için önemli zorluklar yaratır. Görüntüler çok karanlık ve çok gürültülü olur, bu da geleneksel geliştirme tekniklerinin uygulanmasını neredeyse imkansız hale getirir. Son zamanlarda, öğrenme temelli yaklaşımlar bu problem için çok umut verici sonuçlar vermiştir, çünkü daha iyi kaliteyi sağlamak için bu yöntemlerin için ifade gücü yüksektir. Bu yöntemlerden motive olarak, bu tez çalışmasında, iyileştirme performansını artırmak ve aşırı karanlık ham görüntülerden daha keskin ve daha doğru RGB görüntüler elde etmek için seri çekimden yararlanmayı hedefliyoruz. Önerilen yapının bel kemiği, aşamalı olarak yüksek kaliteli çıktılar üreten yeni bir kabadan inceye ağ mimarisidir. Kaba ağ, daha sonra ince ölçekli ayrıntıları ve gerçekçi dokuları kurtarmak için ince ağa beslenen düşük çözünürlüklü bir ham görüntü öngörür. Gürültü seviyesini daha da azaltmak ve renk doğruluğunu artırmak için, bu ağı, giriş olarak seri çekilmiş düşük ışıklı görüntüler alan ve nitelik düzeyinde birden fazla görüntüden bilgi birleştirmesi yapan permütasyon değişmez bir yapıya genişletiyoruz. Deneylerimiz, yaklaşımımızın en son yöntemlerden daha detaylı ve çok daha yüksek kalitede görüntüler ürettiğini ve görsel olarak daha hoş sonuçlar verdiğini göstermektedir.

Anahtar Kelimeler: Hesaplamalı Fotoğrafçılık, Düşük Işıklı Görüntü İşleme, Görüntülerden Gürültü Arındırma, Basamaklı Fotoğrafçılık

ACKNOWLEDGEMENTS

I would like to express my greatest appreciation to my supervisors Assoc. Prof. Dr. Erkut Erdem and Assoc. Prof. Dr. Aykut Erdem. They always encouraged me and pushed me to challenge myself. Their experience and knowledge helped me a lot to shape the ideas presented in this thesis. Their belief in this study was one of the essential components that motivated me for achieving better. Without their guidance, it would have been burdensome to write this thesis.

I also would like to thank my thesis committee members Prof. Dr. Pınar Duygulu Şahin, Prof. Dr. Ahmet Oğuz Akyüz, Assoc. Prof. Dr. Nazlı İkizler Cinbiş and Assoc. Prof. Dr. Tolga Çukur for reviewing this thesis and giving helpful feedback.

I am grateful to my parents for supporting me unconditionally throughout my entire life. It is impossible to thank them enough for their patience and care. I thank my brother Talha Karadeniz for leading the way for me with his experience and teaching me about programming. I owe a debt of gratitude to Mehmet Fatih Karadeniz for being more than a brother in my whole life and helping me with all aspects of life.

I would like to thank Mehmet Emin Gülşen who has been one of the co-founders of our startup together with Mehmet Fatih Karadeniz. We always dreamed beyond the limits and I am sure that we will be proud of each other and what we achieve in the future as today. I also would like thank to Mustafa Çağrı Güven, Tevfik Aktay, Çağatay Yıldız for their intimate friendships.

Finally, words are not enough to express my love for İrem Aybala Gülen. From the beginning of our first met to today, she always gave me courage and inspiration. We have lived the toughest and best moments of life together. She never let me fall and motivated me by reminding me what I can achieve at the times where I doubted myself. I am happy to walk with her on the long and difficult journey of life and I am lucky to have her as an indispensable part of my life.

GENİŞLETİLMİŞ ÖZET

Düşük ışık koşullarında fotoğraf çekmenin çeşitli zorlukları vardır. Temel zorluk, kamera sensörleri tarafından ölçülen sinyal seviyesinin genellikle ölçümlerdeki gürültüden çok daha düşük olmasıdır [1]. Gürültüye neden olan temel faktörler, kamera merceğine giren foton sayısındaki değişimler ve sinyal okunurken meydana gelen sensör tabanlı ölçüm hatalarıdır [2, 3]. Buna ek olarak, düşük ışıklı bir görüntüde bulunan gürültü, ince ölçekli yapıları ve renk dengesi gibi çeşitli görüntü özelliklerini de etkileyerek görüntü kalitesini daha da düşürür.

Düşük ışık koşullarında parlak fotoğraflar çekmek için kamera merceğinin açıklığının genişletilmesi, pozlama süresinin uzatılması veya kamera flaşı kullanılması gibi doğrudan yaklaşımlar bulunur [1, 4]. Bununla birlikte, bu yöntemler, her birinin farklı dezavantajları olduğundan, sorunu tamamen çözmez. Örneğin, açıklığın artırılması donanım kısıtlamaları ile sınırlıdır veya flaş kullanıldığında, fotoğraf makinesine yakın nesneler uzaktaki nesnelerden daha fazla aydınlık gözükür [5]. Uzun pozlama süreleriyle çekilen görüntüler, kameradaki sarsıntı veya sahnedeki nesne hareketleri nedeniyle istenmeyen görüntü bulanıklığına neden olabilir [6]. Bu nedenle, literatürde, geleneksel gürültü giderme ve görüntü iyileştirme yöntemlerinden öğrenmeye dayalı yaklaşımlara kadar düşük ışıklı görüntülerin kalitesini artırmaya çalışan çok çeşitli çalışmalar yapılmıştır.

Görüntü arındırılması, gürültülü bir görüntüden temiz bir görüntüyü geri getirmenin amaçlandığı görüntü işlemedeki klasik problemlerden biridir. Yıllar boyunca görüntüleri arındırmak için çeşitli yöntemler önerilmiştir [7–19]. Bu yaklaşımların çoğu, gürültü gidermek üzerine bir model geliştirmek için Gauss gürültüsü olan görüntülere dayanır. Son zamanlarda, gerçek gürültüler ile başa çıkabilen derin öğrenme tabanlı yöntemler önerilmiştir [3, 20]. Bununla birlikte, bu yaklaşımlar, standart gürültülü bir görüntüden daha zor olan aşırı düşük ışıklı görüntüler için özelleşmemiştir. Görüntü iyileştirme, son birkaç yıl içinde derin öğrenme ile büyük ilerleme kaydeden başka bir aktif araştırma alanıdır. [21–27]. Genellikle, bu yöntemler düşük dinamik aralıklı (LDR) görüntüler ile çalışır ve bu nedenle, kamera işleme hattında biriken hatalar nedeniyle performansları da sınırlıdır. LDR görüntülerle

karşılaştırıldığında, doğrudan fotoğraf makinesinden gelen ham görüntüler, daha fazla bilgi içerdikleri ve minimum işlendikleri için aşırı düşük ışıklı görüntüleri iyileştirmek için daha uygundur.

Aşırı karanlık görüntüleri iyileştirme bağlamında, Seeing in the Dark (SID) [28], standart kamera hattının yerini alacak ilk öğrenme temelli girişimdir ve bir evrişimli sinir ağ (CNN) modeli ile tek bir düşük ışıklı ham görüntüden iyileştirilmiş bir RGB görüntüsü üretir. Bu amaçla, yazarlar kısa pozlama ile karanlık ham fotoğrafları ve bu fotoğraflara karşılık gelen uzun pozlama referans fotoğrafları çekerek veri kümesini topladılar. Bu yöntem, kullanılan mimarideki veya objektif fonksiyonundaki bazı değişikliklerle Maharjan ve diğerleri [29] ve Zamir ve diğerleri [30] tarafından daha da geliştirilmiştir. Benzer şekilde, bu çalışmamızda, tek görüntü geliştirme için yeni bir çok ölçekli mimari geliştiriyoruz ve görsel ve piksel bazlı amaç fonksiyonlarını birleştirerek farklı bir objektif fonksiyonu kullanıyoruz. Ek olarak, önceki yöntemler tek bir karanlık ham görüntüden bir RGB görüntüsü elde ederken, bu çalışmada sahne ile ilgili birden fazla gözlemi birleştirerek sonuçların iyileştirilip iyileştirilemeyeceğini araştırıyoruz.

Basamaklama yöntemi, fotoğrafçılıkta aynı sahnenin birkaç fotoğrafını hızla çekmeye dayanan, iyi bilinen bir tekniktir. Bu çekimler genellikle sahnenin özelliklerini farklı şekilde yakalayan pozlama gibi bazı kamera ayarları açısından birbirinden farklıdır ve böylece yüksek dinamik aralıklı (HDR) bir görüntü oluşturmak gibi uygulamalar için kullanılabilir. Pozlama basamaklamaya benzer bir teknik, seri çekimdeki her kareyi sabit bir pozlama ile çekmektir [4]. Düşük ışık altında sürekli kısa pozlama ile çekildiğinde, bu görüntüler aynı sahnenin farklı karanlık, gürültülü gerçekleşmelerini temsil eder. Doğal olarak, bize tek bir karanlık görüntü ile karşılaştırıldığında sahne hakkında daha fazla gözlem sağlıyorlar. Bu görüntülerin ortalamasının alınması gürültüyü azaltırken, sonuçlar her zaman tatmin edici değildir. Bu nedenle, seri çekimdeki geçici pikselleri birleştirmek için farklı teknikler önerilmiştir [1, 4, 31–36]. Bu yaklaşımlar arasında, [34–36], seri çekim görüntülerini işlemek için öğrenme tabanlı yöntemler kullanır. Bu çalışmalarda, seri görüntüler CNN'ye kanallar yoluyla birleştirilerek veya tekrarlayan bir şekilde beslenir. Bizim durumumuzda, radikal olarak farklı bir yaklaşım öneriyoruz ve bu seri çekim görüntülerini permütasyon

değişmez bir şekilde işlemenin basit ama daha etkili bir yaklaşım olduğunu gösteriyoruz. Permütasyon değişmez yapılarda, seri çekim görüntülerinin sırası çıktıyı etkilemez ve buna bağlı olarak daha doğru bir çıktı elde edilebilir. Önceki çalışmalar [28–30, 36], kayda değer ilerleme göstermelerine rağmen, aşırı düzleştirme ve renk yanlışlıkları gibi yapaylıklardan muzdaripler ve görüntüdeki ince ölçekli ayrıntıları kurtaramıyorlar.

Özetle, bu eksiklikleri hafifletmek için, bu çalışmada, bir sahnenin aşırı düşük ışıklı seri çekim ham görüntülerini girdi olarak alıp iyileştirilmiş bir RGB görüntü oluşturan öğrenme tabanlı bir yapı öneriyoruz. Bunun için, seri çekim karanlık görüntüleri eşzamanlı olarak işleyebilen bir kabadan inceye permütasyon değişmez ağ mimarisi geliştiriyoruz.

CONTENTS

Page

TABLES

FIGURES

1. INTRODUCTION

Low-light image enhancement is one of the important tasks in image processing and computer vision where the aim is to improve the quality of the images taken in the dark environments. There are two main reasons for why low light image enhancement is important. First, generating higher quality images by digital cameras leads to higher end-user satisfaction. Hence, a number of camera and mobile-device companies recently developed different imaging pipelines to address low-light image enhancement problem [1]. Second, low-light conditions may introduce certain challenges for high-level vision tasks. Because of this, enhancing dark images is not only valuable for obtaining visually satisfying images but can also improve the performance of other approaches that are effective for images captured in ideal illumination environments.

Darkness of a scene is measured in terms of lumens per meter squared (lux). Descriptions of the lux levels provided by Levoy [37] can be seen in Table 1.1., where scenes below 3 lux can be considered dark. In addition to the illumination conditions of a scene, camera parameters also play an important role when capturing images in low-light. These parameters are mainly ISO, aperture and shutter speed. As the shutter speed decreases, the amount of the light entering the lenses of the camera increases which results in brighter images. Similarly, aperture also affects the amount of light entering the lens. Another important camera parameter is the ISO parameter which controls the sensitivity of the camera to the light. When ISO increases, the resulting image becomes brighter since the camera becomes more sensitive to the light.

Capturing images in low-light conditions is challenging – the main difficulty being that the level of the signal measured by the camera sensors is generally much lower than the noise in the measurements [1]. The fundamental factors causing the noise are the variations in the number of photons entering the camera lens and the sensor-based measurement errors occurred when reading the signal [2, 3]. In addition, noise present in a low-light image also

Lux	Description
30,000	Sidewalk lit by direct sunlight
1,000	Sidewalk on an overcast day
300	Typical office lighting
20	Restaurant with atmospheric lighting
3	Sidewalk lit by street lamps
1	Limit of reading a newspaper
0.6	Sidewalk lit by full moon
0.1	Wouldn't walk through the house without the flashlight

Table 1.1. Descriptions of lux levels [37]

affects various image characteristics such as fine-scale structures and color balance, further degrading the image quality.

Direct approaches for capturing bright photos in low light conditions include widening the aperture of the camera lens, lengthening the exposure time, or using camera flash [1, 4]. These methods, however, do not solve the problem completely as each of these hacks has its own drawbacks. Opening the aperture is limited by the hardware constraints, and when the camera flash is used, the objects closer to the camera are brightened more than the objects or the scene elements that are far away [5]. Images captured with long exposure times might have unwanted image blur due to camera shake or object movements in the scene [6]. Hence, in the literature, there have been a wide range of studies which try to improve the quality of low-light images, ranging from traditional denoising and enhancement methods to learning-based approaches.

Image denoising is one of the classical problems in image processing, where the aim is to restore a clean image from a noisy image. Several methods have been proposed over the years to denoise images [7–19]. Most of these approaches rely on the images with Gaussian noise for developing a denoising model. Recently, deep learning-based methods that can deal with real image noise have been proposed [3, 20]. However, these approaches are not specialized to extremely low-light images which are harder to restore than a standard noisy image. Image enhancement is another active field of research, which has seen tremendous

progress in the past few years with deep learning [21–27]. Usually, these methods work with low dynamic range (LDR) input images and hence, their performance is also limited due to the errors accumulated in the camera processing pipeline. When compared to LDR images, raw images straight from the camera are more suitable to use for enhancing extremely low-light images since they contain more information and are processed minimally.

In the context of enhancing extremely dark images, See-in-the-Dark (SID) [28] is the first learning-based attempt to replace the standard camera pipeline, training a convolutional neural network (CNN) model to produce an enhanced RGB image from a single raw low-light image. For this purpose, the authors collected a dataset of short-exposure, dark raw photos and their corresponding long-exposure references. Their method is further improved by Maharjan et al. [29] and Zamir et al. [30] with some changes in the CNN architecture and the objective functions utilized in training. In a similar fashion, in our study, we develop a new multi-scale architecture for single image enhancement and use a different objective by combining contextual and pixel-wise losses. While the previous methods obtain an RGB image from a single dark raw image, we further explore whether the results can be improved by integrating multiple observations regarding the scene.

Bracketing is a well-known technique in photography that relies on rapidly taking several shots of the same scene. These shots usually differ from each other in terms of some camera settings, e.g. exposure, which capture characteristics of the scene differently, and thus they can be used for applications like constructing a high dynamic range (HDR) image. A technique similar to exposure bracketing is shooting each frame in the burst sequence with a constant exposure [4]. To our interest, when shot with a constant short exposure under low-light, these images represent different dark, noisy realizations of the same scene. Naturally, they provide us multiple observations about the scene when compared to a single dark image. While simply averaging these images reduces noise, results are not always satisfactory. For this reason, different techniques are introduced to merge the temporal pixels in the burst sequence [1, 4, 31–36]. Among these approaches, [34–36] use learning-based methods to process burst images. In these studies, burst images are fed to a CNN either by concatenating through channels or in a recurrent fashion. In our case, we propose a radically different

approach and show that processing these burst images in a permutation invariant manner is a simple yet more effective approach. The order of burst images does not affect the output, and accordingly a more accurate output can be obtained. In Fig. 1.1., we present the results of the aforementioned extremely low-light image enhancement models along with our results. The multiple image enhancement models, which either employ burst imagery or integrate ensemble of enhanced images, give superior results than their single image counterparts, yet they still suffer from artifacts such as over-smoothing, and fail to recover fine-scale details in the image. Despite the remarkable progress of previous studies [28–30, 36], this example image demonstrates that there is still large room for improvement, regarding various issues such as unwanted blur, noise and color inaccuracies in the end results – especially for the input images which are extremely dark.

In a nutshell, to alleviate these shortcomings, in this thesis, we propose a learning-based framework that takes a burst of extremely low-light raw images of a scene as input and generates an enhanced RGB image. In particular, we develop a coarse-to-fine network architecture which allows for simultaneous processing of a burst of dark raw images as input to obtain a high quality RGB image.

Our main contributions are summarized as follows:

- We introduce a multi-scale deep architecture for image enhancement under extremely dark lighting conditions, which consists of a coarse-scale network and a fine-scale network.
- We further extend our coarse-to-fine architecture to design a novel permutation invariant CNN model that predicts an enhanced RGB image by integrating features from a burst of images of a dark scene.
- Our experiments demonstrate that our approach outputs RGB images with less noise and sharper edge details than those of the state-of-the-art methods. These are validated quantitatively based on several quality measures in both single-frame and burst settings.

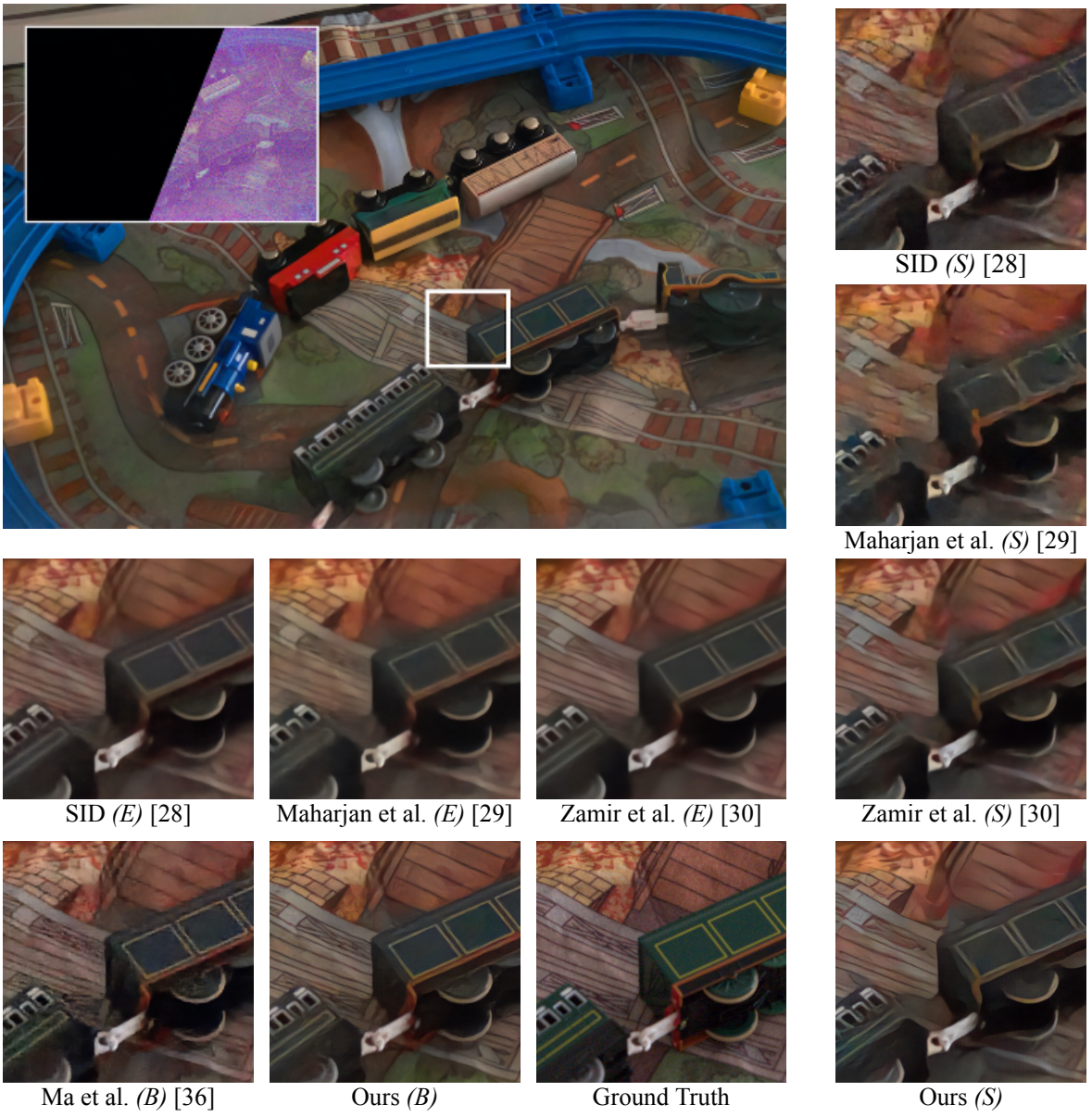


Figure 1.1. A sample result obtained with our proposed burst-based extremely low-light image enhancement method. The standard camera output and its scaled version are shown at the top left corner. For comparison, the zoomed-in details from the outputs produced by the existing approaches are given in the subfigures. The results of the single image enhancement models, denoted with *(S)*, are shown on the right. The results of the multiple image enhancement methods are presented at the bottom, with *(B)* denoting the burst and *(E)* indicating the ensemble models. Our single image model recovers finer-scale details much better than its state-of-the-art counterparts. Moreover, our burst model gives perceptually the most satisfactory result, compared to all the other methods.

2. RELATED WORK

Low-light images show different characteristics due to the lighting conditions of the environments, and the noise and/or motion blur they contain. In general, the approaches for low-light image processing can be divided into two groups, with respect to the darkness levels of the input images: (i) low-light image enhancement, and (ii) extremely low-light image enhancement. Generic low-light image enhancement methods refer to the approaches that restore the perceptual quality of images taken under poor illumination conditions, which suffer from low visibility. Enhancement models for extremely low-light images, on the other hand, deal with images captured under more severe conditions, which cannot be handled by the first group of works. In particular, the darkness of an image is directly related to the illuminance of a scene, which is measured in terms of lumens per meter squared (lux). In this sense, extremely low light images denote short exposure images (usually between 1/30 and 1/10 sec exposure) that are taken in 0.2-5 lux outdoor or 0.03-0.3 lux indoor scenes.

In this thesis, we explore the use of burst photography for enhancing extremely dark images. Since extremely low-light images contain severe noise, our work is also related to generic image denoising and burst photography. Hence, in this chapter, we provide a brief review of image denoising, low-light image enhancement, extremely low-light image enhancement and burst photography methods proposed in recent years. A taxonomy is given in Fig. 2.1. with accompanying references.

2.1. Image Denoising

Image denoising is a fundamental problem in computer vision that deals with removing noise from an image [40, 41]. Traditionally, methods that exploit the non-local self-similarity prior [7–9], sparsity [10, 11] and image gradients [12] have been widely used for image denoising. Recently, various deep learning approaches have been proposed for both non-blind Gaussian denoising [13, 14] and blind Gaussian denoising [15, 16], which involve training denoising models under known and unknown noise levels, respectively. Lately, researchers proposed

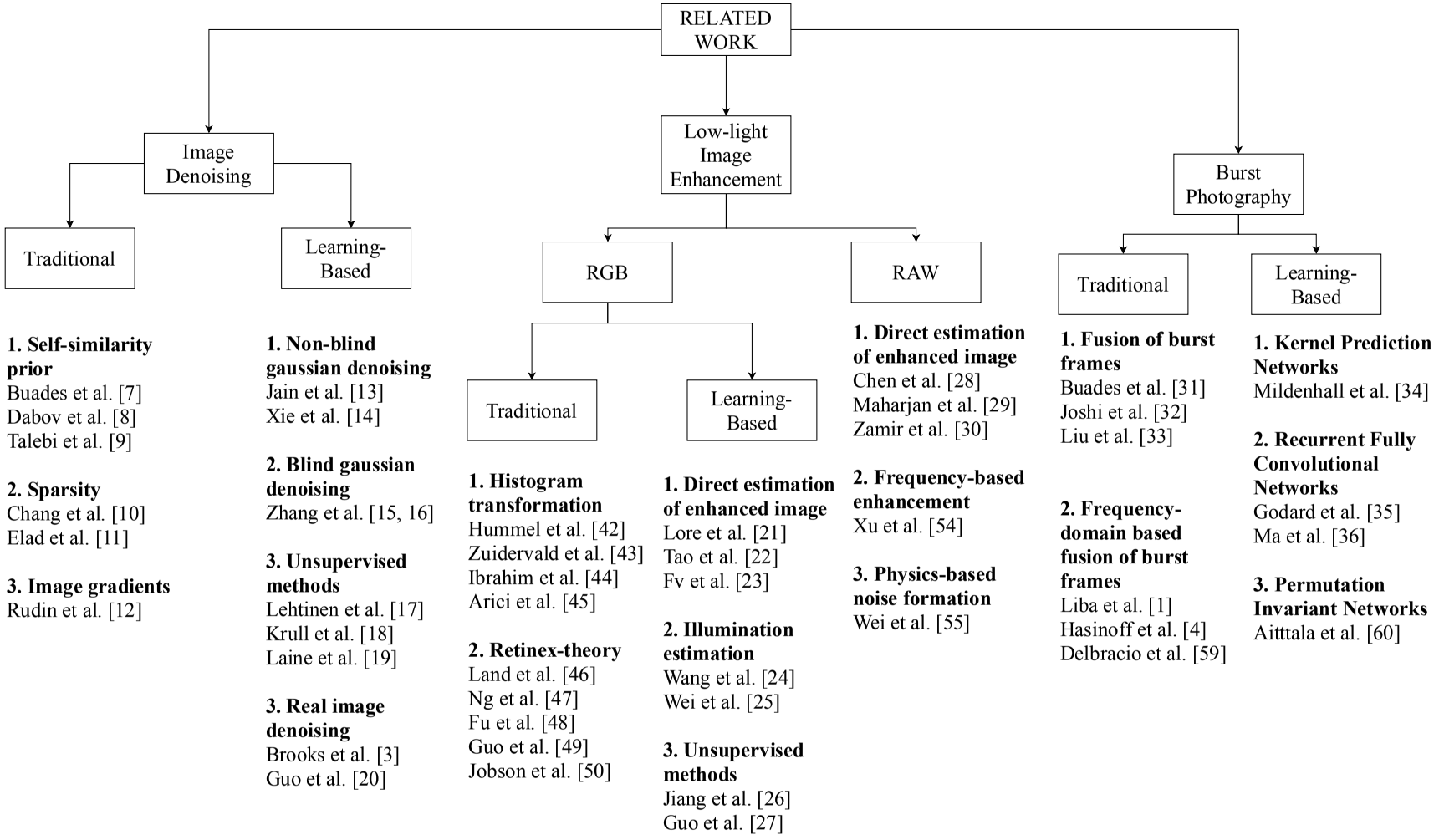


Figure 2.1. Related work for extremely low-light image enhancement

unsupervised deep denoising models [17–19] that do not use any clean ground truth data during training. Although most of these existing denoising models focus on additive white Gaussian noise, this noise model falls short when the real-life images are considered. Hence, the recent trend in image denoising is to develop models that are trained with real-world noisy data [3, 20] and that can generalize much better than the models which consider additive white Gaussian noise. While these aforementioned recent methods give fairly good results most of the time, they are not well-suited to extremely dark images as they suffer from severe noise and color degradation, as shown in Fig. 2.2.

(a) Dark (b) Traditional

(c) Traditional + BM3D denoising (d) Long exposure

Figure 2.2. For an extremely dark image displayed in (a), the traditional camera pipeline produces a highly noisy image with severe color degradation, as shown in (b). Moreover, as demonstrated in (c), the state-of-the-art denoising methods cannot handle these challenges and give unsatisfactory results. Extremely low-light image enhancement methods, on the other hand, aim for generating an output close to a long-exposure image, like the one given in (d).

2.2. Low-Light Image Enhancement

Generic approaches that can be used for low-light image enhancement can be divided into three groups: (i) traditional contrast enhancement methods, (ii) techniques based on Retinex-theory, and (iii) learning-based approaches. Most well-known methods for contrast enhancement include histogram equalization based approaches that apply transformations to image histograms [42–45]. Motivated by human color perception, Retinex-theory based approaches decompose the images into illumination and reflectance components, and take into account these components while enhancing the images [46–50]. On the other hand, learning-based methods mostly include discriminative methods based on sparse autoencoders [21] and CNNs that either directly estimate an enhanced image [22, 23] or extract an illumination map [24, 25]. Recently, researchers suggested some unsupervised models which employ adversarial losses for enhancement [26] or CNNs for illumination curve estimation [27].

These low-light image enhancement methods provide good results under certain conditions. However, they fail to deal with the full extent of the challenges in imaging under extremely dark conditions. These enhancement models mainly accept LDR images generated by the standard camera pipeline. Transforming raw images to LDR images introduces some information loss in the measurements which complicates the enhancement process. Hence, these low-light image enhancement models are favorable only when the input images are partly dark and do not exhibit serious color degradation and severe noise.

2.3. Extremely Low-Light Image Enhancement

As discussed in the introduction, enhancing extremely dark images was introduced as a challenging image enhancement task by Chen et al. in [28], and the See-in-the-Dark (SID) model proposed therein is the first model that specifically aims for solving this task. This approach processes a raw image captured under very poor illumination condition with a U-Net [51] like architecture. Training of the model is carried out on a dataset of paired short and long-exposure images by taking into account a pixel-wise (L_1) loss.

Very recently, there have been a few attempts to further improve the performance of SID. For instance, Maharjan et al. [29] have proposed to use residual learning to boost the final image quality. Zamir et al. [30] have used a hybrid loss function which is a combination of pixel-wise and multi-scale structural similarity (MS-SSIM) losses and a perceptual loss [52, 53], which is defined by the absolute difference of the features extracted by a deep network. Interestingly, in [36], Ma et al. have developed an enhancement model for extremely low-light images, which employs recurrent convolutional neural networks to obtain a high quality result from a burst of input images. Although these studies demonstrate significant progress in enhancing extremely low-light images, they cannot fully deal with the challenges of the dark scenes. As presented in Fig. 2.3., the images enhanced by these approaches may suffer from artifacts such as over-smoothing and color bleeding. Moreover, the existing models do not recover texture and fine details such as thin structures successfully.

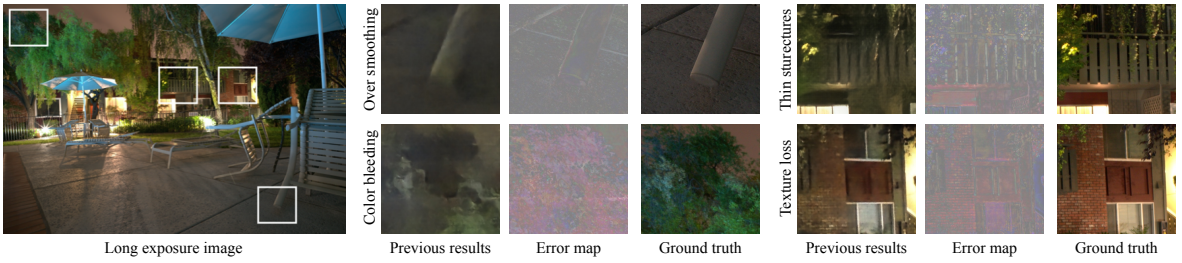


Figure 2.3. Common failure cases for the state-of-the-art extremely low-light image enhancement methods. Subfigures show some cropped images from the results of the existing models together with the corresponding error and the ground truth images, demonstrating that these models suffer from over-smoothing and color bleeding artifacts and fail to properly recover thin structures and textured regions.

Xu et al. proposed a frequency-based enhancement and decomposition method for low-light image enhancement [54]. They developed a network that learns to recover image objects in the low-frequency layer and then enhances the high-frequency details on the recovered objects. Wei et al. suggested a noise formation model to synthesize realistic noisy images that can match the quality of real data under extreme low-light conditions [55]. They additionally presented a noise parameter calibration method that can adapt their model to a given camera.

As will be discussed in the next chapter, different from the aforementioned methods, we alternatively propose a multi-scale approach which uses a novel coarse-to-fine architecture

that better handles the extremely low-light images by giving much sharper and more vivid colors. In addition, we use a combination of the L_1 pixel loss and the recently proposed contextual loss function which maintains the image statistics better [56]. Moreover, for our burst model, we employ a set-based permutation invariant architecture that jointly processes low-light input images in an orderless manner, giving perceptually plausible and high quality results.

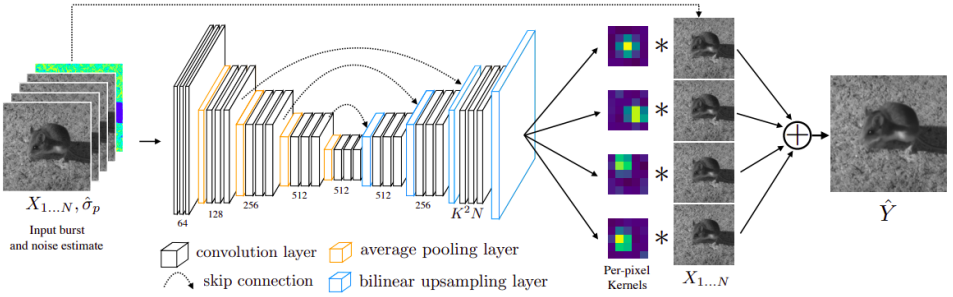
There are also some recent efforts to extend the aforementioned image enhancement models to videos by additionally taking into account temporal consistencies. For instance, Chen et al. [57] extended their SID model to videos by training a Siamese network on static raw videos. Similarly, Jiang and Zheng [58] proposed a U-Net like architecture containing 3D convolution layers for the same purpose. These models are out of focus of this study as they require training with dark videos not images, but are mentioned here only for completeness.

2.4. Burst Photography

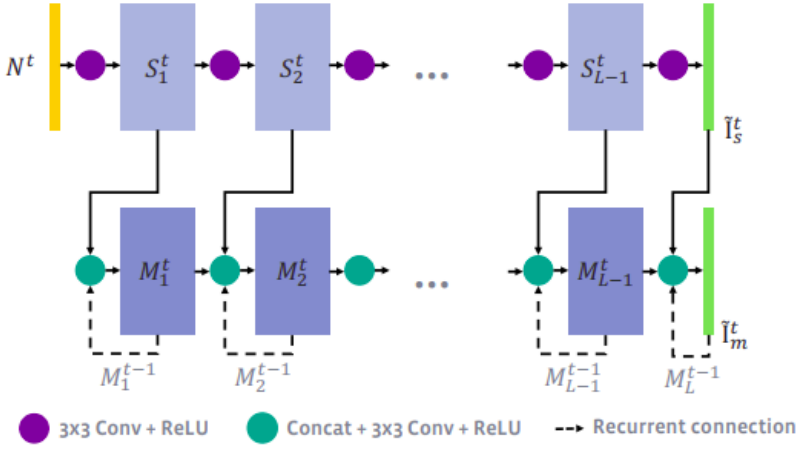
Burst photography refers to the process of capturing a sequence of images each spaced a few milliseconds apart and subsequently integrating them to obtain a higher-quality image. For instance, the most intuitive way to produce a noise-free image is to capture a burst of images and apply simple averaging. Yet, this strategy gives unsatisfactory results in practice due to moving objects and/or a moving camera. Hence, a variety of more complicated methods were introduced to combine the information from multiple images in a more effective manner. Buades et al. proposed to apply standard averaging only for the aligned pixels and utilize the state-of-the art denoising methods for the remaining pixels [31]. Joshi et al. developed a method that weights the pixels with respect to their sharpness levels by using Laplacian convolution [32] and accordingly utilizes these weights in obtaining higher quality images. Liu et al. proposed to fuse the consistent pixels with an optimal linear estimator [33]. Moreover, some researchers suggested to employ the information encoded in the frequency-domain for temporal fusion [1, 4, 59]. Recently, more sophisticated approaches have been proposed for denoising such as Kernel Prediction Networks [34], Recurrent Fully Convolutional Networks

[35], and Permutation Invariant Networks [60], which process a burst of noisy and blurred images through deep CNN architectures, as shown in Fig. 2.4.

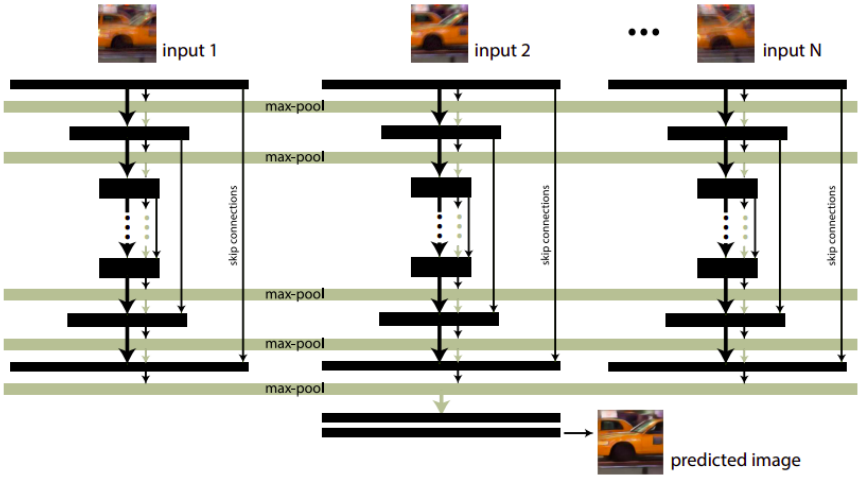
These aforementioned models do not cope with the challenges of extremely dark images – with the exception of Liba et al. [1] and Hasinoff et al. [4] where the authors rely on hand-crafted strategies. As mentioned in the previous section, the only work that focuses on learning-based burst imagery in the extremely low-light conditions is the work by Ma et al. [36]. In this work, the authors utilized a recurrent convolutional neural network architecture, similar to the one in [35], to enhance a burst of raw low-light images. In our work, specifically motivated by these recent burst photography approaches, we develop a set-based permutation invariant CNN architecture that can be used to obtain a high quality image from a burst of extremely dark images. In particular, our network jointly processes the burst frames in an orderless manner, as compared to the recurrent model in [36] which processes each frame sequentially.



(a) Kernel Prediction Networks [34]



(b) Recurrent Convolutional Networks [35]



(c) Permutation Invariant Networks [60]

Figure 2.4. Different CNN architectures for burst image processing.

3. PROPOSED APPROACH

Table 3.1. summarizes the notations used throughout this study. Our aim is to learn a mapping from the domain of raw low-light images to the domain of long-exposure RGB images. To achieve this, we first propose a single-frame coarse-to-fine model and then extend it to a set-based formulation to process a burst of images. The details of our networks are illustrated in Fig. 3.1.

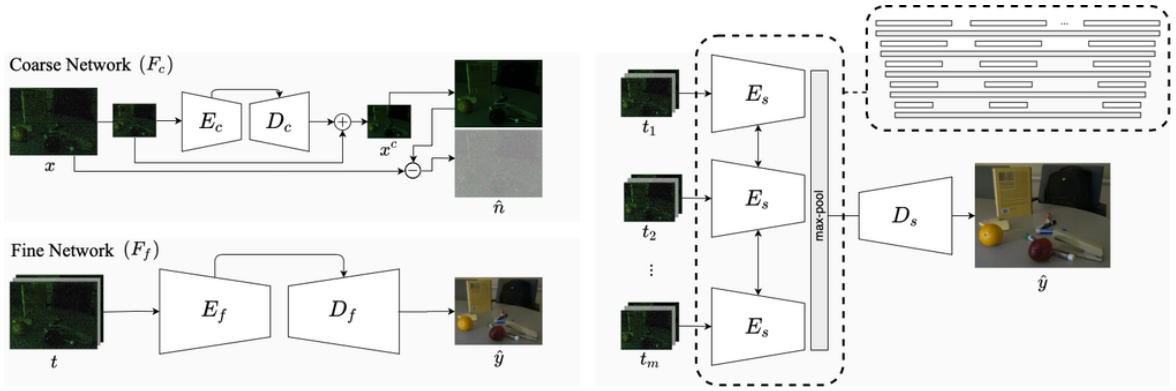
Preserving structures and colors in dark images is difficult because of the extreme noise present in these images. Although one can increase the complexity of a CNN to develop a model that can adapt to various levels of noise and darkness, increasing the number of parameters does not always increase the generalization capabilities of the network. Moreover, we usually need large effective receptive fields so that we obtain images with globally correct colors and edge details. Then, we can use another network to increase the details in these images. Furthermore, if we knew how dark an image is before feeding it to our network, we could use this information for guiding the network about the noise level. Thus, developing a multi-scale network enables us to address these problems.

Multiple observations about a scene increase the possibility of recovering a signal under extreme darkness. We can use bracketing technique to capture multiple frames of a scene. Although we capture the same scene multiple times, each of these frames differs from each other in terms of noise. A naive approach to integrating these frames into a single frame is frame averaging. However, the resulting quality is not always better than the quality of an image obtained via a CNN whose input is a single dark raw image. To solve this problem, we propose to use a network that can process multiple frames simultaneously. Hence, we develop a set-based network which can extract features from each frame and merge these features to obtain an enhanced image.

To recover fine-grained details from dark images, we propose to employ a two-step coarse-to-fine training procedure. Similar strategies have been proven very effective in various other tasks such as deblurring [61] and image synthesis [62]. Different than those approaches, our

$x_1, x_2, \ldots, x_m$	Burst of raw low-light input images
y, $\hat{y}$	Reference and predicted long-exposure RGB images
$F_c(\cdot), F_f(\cdot), F_{\text{s}}(\cdot)$	Coarse, fine and set-based networks
$x_1^c, x_2^c, \ldots, x_m^c$	Raw, low-res outputs of the coarse network
$\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_m$	Noise approximations for $x_1, x_2, \ldots, x_m$
$t_1, t_2, \ldots, t_m$	Tensors containing raw inputs, upsampled coarse outputs and noise approximations
$R_d(\cdot), R_u(\cdot)$	Downsampling and upsampling functions

Table 3.1. The notations used throughout this study.



(a) Coarse-to-fine network (b) Set-based network

Figure 3.1. Network architectures of the proposed (a) single-frame coarse-to-fine model, and (b) set-based burst model.

coarse network outputs a raw (denoised) image. This helps us to decouple the problem of learning the mapping between the raw domain and the RGB domain. Some recent denoising methods use the noise level as an additional input channel [3, 34]. Predicting the coarse outputs in the raw domain also allows us to compute the approximate noise in the input.

In our proposed framework, the raw low-light input image is first downsampled by a factor of two and then fed to our coarse network. The coarse network, which is illustrated in Fig. 3.1.(a), is trained on downsampled data and produces denoised and enhanced outputs in low-resolution

$$x^c = F_c(R_d(x)). \tag{1}$$

We utilize the output of the coarse network not just for guidance in assisting the fine network but also in approximating the noise by computing the difference between the upsampled coarse prediction and the raw low-light input, as:

$$\hat{n} = x - R_u(x^c) \tag{2}$$

The fine network takes the concatenation of the low-light raw input image, the output from the coarse network and the noise approximation as inputs and processes them to generate the final RGB output:

$$\hat{y} = F_f(t), \quad t = (x, \hat{n}, R_u(x^c)) \tag{3}$$

Both our coarse and fine networks follow a U-Net like encoder-decoder architecture. In the encoder, they contain 10 convolution layers where the number of filters is doubled and the resolution is halved after every 2 convolution layers, with the initial number of filters is set to 32. In the decoder, they include deconvolution layers which are concatenated with earlier corresponding convolution layers through skip connections. As shown in Fig. 3.2., the coarse network gives a fairly good enhancement result for a given extremely low-light image containing severe noise and color degradation. The fine network further improves the color accuracy and the details of the result of the coarse network, producing a higher quality image.

3.1. Set-Based Extension to Burst Images

Recently, there have been some attempts to study the invariance and equivariance properties of neural networks [63–65]. Interestingly, Zaheer et al. provided a generic algorithm to train neural networks that operate on sets via a simple parameter sharing scheme [66], which allows for information exchange with a commutative operation. Based on this idea, Aittala and Durand proposed a permutation invariant CNN model for burst image deblurring [60]. In a similar vein, in this thesis, we develop a permutation invariant CNN architecture but with a much lower computational cost by using multiple encoders and a single decoder.

(a) Traditional (b) Coarse (c) Fine (d) Burst

Figure 3.2. An example night photo captured with 0.1 sec exposure and its enhanced versions by the proposed coarse, fine and burst networks. As the cropped images demonstrate, the fine network enhances both the color and the details of the coarse result. The burst network produces even much sharper and perceptually more pleasing output.

We extend our coarse-to-fine model to a novel permutation invariant CNN architecture which takes multiple images of the scene as input and predicts an enhanced image. In particular, first, low-resolution coarse outputs are obtained for each frame x_i in the burst sequence, using our coarse network:

$$x_i^c = F_c(R_d(x_i)) \tag{4}$$

In addition, we compute an approximate noise component n_i for each frame, as

$$\hat{n}_i = x_i - R_u(x_i^c) \ . \tag{5}$$

Finally, our set-based network accepts a set of tensors $\{t_i\}$ as input, each instance $t_i = (x_i, \hat{n}_i, R_u(x_i^c))$ corresponding to the concatenation of one of raw burst images x_i, its noise approximation $\hat{n}_i$ and the upsampled version of the coarse prediction $R_u(x_i^c)$, and produces the final RGB output:

$$\hat{y} = F_s\left(\{t_1, \ldots, t_m\}\right) \ . \tag{6}$$

In the above equation, F_s represents our permutation invariant CNN, which has m convolutional subnetworks which allow for information exchange between the features of burst frames. This is achieved by using a max-pooling over the set of burst features after each

convolution layer in the encoder part of the network. Then, in the decoder part, instead of concatenating the deconvolution features with the corresponding earlier features, we concatenate them with the corresponding global max-pooled features computed in the encoder part. Hence, without even changing the parameter size, we integrate the advantage of multiple observations to the network. As Fig. 3.2. demonstrates, processing multiple dark images via the proposed burst network significantly improves the quality of the end result. Our burst network produces perceptually more pleasing and sharper results than our fine network and especially recovers the fine details and the texture much better.

3.2. Losses

To train our networks, we tested combining a pixel-wise loss (L_1) with two alternative featurewise losses, namely perceptual loss (L_{P}) [52, 53] and contextual loss (L_{CX}) [56, 67].

Pixel-wise Loss. Since we want output images to be close to long exposure images, we compute the pixel-wise reconstruction error as the objective function. For the image denoising and demosaicking task, L_1 performs better than L_2 as large errors are over-penalized by L_2, especially in smooth regions [68, 69]. Furthermore, it is also shown in [68] that even when the output quality is measured by L_2 metric, a network trained with an L_1 loss can result in lower L_2 error because of their convergence properties. Therefore, we preferred to use the L_1 loss between the network output and the ground truth long-exposure image, given as

$$L_1(y, \hat{y}) = \|y - \hat{y}\|_1 . \tag{7}$$

Perceptual Loss. Using only pixel-wise losses may result in blurry images [52]. Although humans may tolerate small transformations of images, they are sensitive to perturbations in structure information in the images such as sharpness. Similarly, feature maps of CNNs trained for core vision problems can represent perceptually important properties and are invariant to transformations that preserve the content [52, 70]. Indeed, recent studies on perceptual image metrics [38, 71] suggest that features of CNNs trained for high-level vision

tasks agree with humans when judging the perceptual quality of images. Thus, to measure the distance at a more semantic level, we employ the commonly used perceptual loss [52, 53], which uses high-level features from a pre-trained VGG-19 network [72], defined as

$$L_{\mathrm{P}}(y, \hat{y}, l) = \left\| \phi^l(y) - \phi^l(\hat{y}) \right\|_1 \tag{8}$$

where $\phi^l(\cdot)$ denotes the feature maps at the l-th layer of the network.

Contextual Loss. Assume that the output image of a network is perceptually similar to its ground truth but not spatially aligned. Then, both pixel-wise and perceptual losses may penalize the misalignments despite the perceptual similarity. Contextual loss proposed in [56, 67] aims to measure the similarity of two non-aligned data by comparing sets of features. According to Mechrez et al., by comparing feature distributions, the network also learns to better capture changes in fine-scale details. Specially, contextual loss measures the statistical difference between the feature distributions $\phi^l(y)$ and $\phi^l(\hat{y})$ extracted from y and $\hat{y}$, respectively, and it is defined as

$$L_{\mathrm{CX}}(y, \hat{y}, l) = -\log(\mathrm{CX}(\phi^l(y), \phi^l(\hat{y}))) \tag{9}$$

where the statistical similarity CX is estimated by an approximation of the KL-divergence, as follows.

Let $R = \{r_i\}$ and $S = \{s_j\}$ respectively represent the set of features extracted from a pair of images, with cardinality N, and d_{ij} be the cosine distance between the features r_i and s_j. Then, $\mathrm{CX}(R, S) = \frac{1}{N} \sum_j \max_i \mathrm{CX}_{ij}$ where $\mathrm{CX}_{ij} = w_{ij} / \sum_k w_{ik}$ and $w_{ij} = \exp\left(\frac{1-\tilde{d}_{ij}}{h}\right), \tilde{d}_{ij} = \frac{d_{ij}}{\min_k d_{ik} + \epsilon}$.

Implementation Details. To generate our training data, we extracted 512×512 pixels random patches for each input image and also generated their downsampled versions with half resolution (obtained by bilinear interpolation). Hence, the input patch sizes for the coarse and fine networks are 256×256 and 512×512 pixels, respectively. We first trained the coarse network F_c by using Adam optimizer with a learning rate of 10^{-4} for 2000 epochs and 10^{-5} for

2000 epochs. Then, the fine network F_f was trained with the same hyperparameters without fixing the parameters of the coarse network. Finally, we trained the set-based network F_s for 1000 epochs by initializing its weights from the fine network. During the training of F_s, we randomly chose the number of burst input frames between 1 and 10. We trained both of our models by using a hybrid loss that consists of the pixel-wise L_1 and the contextual L_{CX} loss functions[1]. For the contextual loss, we used `conv3_2` and `conv4_2` layers of the VGG-19 network. We implemented our model with Tensorflow library on an NVIDIA GeForce GTX 1080 Ti GPU. Training our model lasted about 4 days.

[1] In our experiments, we observed that the contextual loss L_{CX} works consistently better than the perceptual loss L_{P}.

4. EXPERIMENTAL EVALUATION

4.1. Dataset

Obtaining long-exposure images is practically difficult but they can serve as ground truth images if the low-light scenes are static. We train and evaluate our models on the SID dataset [28], which consists of short-exposure burst raw images taken under extremely dark indoor (0.2-5 lux) or outdoor (0.03-0.3 lux) scenes. These images are acquired with three different exposure times of 1/10, 1/25 and 1/30 sec, where the corresponding reference images are obtained with 10 seconds or 30 seconds exposures depending on the scene. Specifically, we evaluate the performance of our models on the Sony subset, which contains 161, 36 and 93 distinct burst sequences for training, validation and testing, respectively. The number of burst frames varies from 2 to 10 for each distinct scene. The burst images are totally aligned as they are captured with a tripod. The total number of images in this dataset is 2697, including the burst frames. Moreover, the images are categorized into three groups based on their amplification ratios ($\times$100, $\times$250, $\times$300), measured as the ratio between the exposure times of the dark input image and the long-exposure ground truth. Some example images from the SID dataset are shown in Fig. 4.1.

4.2. Competing Approaches

We compare our models with four state-of-the-art methods, SID [28], Maharjan et al. [29], Ma et al. [36] and Zamir et al. [30]. In our experiments, we used the pre-trained models provided by the authors of [28] and [29], and our implementations of the methods in [36] and [30] as their models are not publicly available. Specifically, for the method of Zamir et al. [30], we trained the U-Net model with the hybrid loss including pixel-wise L_1 and MS-SSIM losses and the perceptual loss L_{P} for 4000 epochs. For the burst-based model by Ma et al. [36], we implemented a recurrent U-Net architecture, where the concatenated features from the previous frame, the single image model and the previous layer are fed to each convolution block of the network. We trained the model for 1000 epochs fixing the

(a) Dark (b) Traditional (c) Long exposure

Figure 4.1. We show example (a) short exposure, dark raw images, (b) scaling and demosaicking dark images via traditional pipeline, (c) long exposure images from the SID dataset.

parameters of the single image network. It is important to note that among these approaches, only the method by Ma et al. [36] processes a burst of images at once. For a fair comparison with the single image models, we also process each burst image independently via each model, take the average of these enhanced outputs as the final result, and additionally report the predictions of these ensemble models.

4.3. Evaluation Metrics

For quantitative evaluation, we employ the popular peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) metrics and also two perceptual image quality metrics, namely learned perceptual image patch similarity (LPIPS) [38] and perceptual image-error assessment through pairwise preference (PieAPP) [71]. These perceptual metrics can be used to quantify the natural distortion of images such as noise and blur as well as CNN-based distortions. In addition, we also utilize perceptual index (PI) [73], a recently proposed no-reference perceptual image quality metric.

Peak Signal-to-Noise Ratio. Signal-to-noise ratio (SNR) is a term that measures the ratio of the level of a signal to the level of noise. It can be defined in dB as the following [74]:

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)^2}{\sum_{x=0}^{M-1} \sum_{y=0}^{N-1} [f(x,y) - \hat{f}(x,y)]^2} \tag{10}$$

where $\hat{f}$ is an approximation of the uncorrupted $M \times N$ image f. Note that SNR depends on the average intensity of the image. By replacing the numerator with the square of the maximum possible value of the image, we obtain

$$\mathrm{PSNR} = 10 \log_{10} \frac{\max(f(x,y))^2}{\frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} [f(x,y) - \hat{f}(x,y)]^2}. \tag{11}$$

The only difference between mean squared error and PSNR is that we can compare images with different dynamic ranges with PSNR [75].

Structural Similarity Index Measure. Structural similarity index mesaure (SSIM) is an alternative for PSNR that also considers the structural information of the scene[76]. SSIM is based on the contrast, luminance and structure properties of images. Let x and y be two images to be compared. Then, luminance of x and y are estimated as the mean intensities μ_x

and μ_y. The luminance comparison function is defined as

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}. \tag{12}$$

Then, mean intensities are subtracted from the images and standard deviations of images are used to define contrast function

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}. \tag{13}$$

Structure comparison is estimated after mean subtraction and variance normalization. It is defined as

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{14}$$

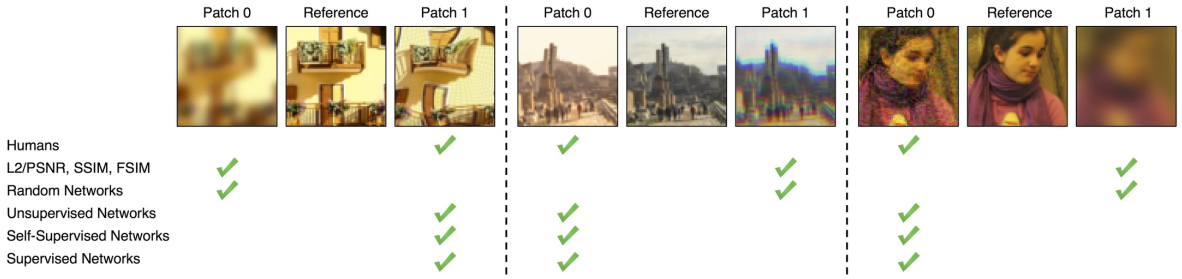
where σ_{xy} is the covariance of x and y. The constants $C_1 = (K_1L)^2$,$C_2 = (K_2L)^2$ and $C_3 = C2/2$ are used to prevent instabilities where $K_1, K_2 << 1$ and L is the dynamic range of images. Finally, the structural similarity index of x and y is defined as a combination of these functions as

$$\mathrm{SSIM}(x,y) = [l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma] \tag{15}$$

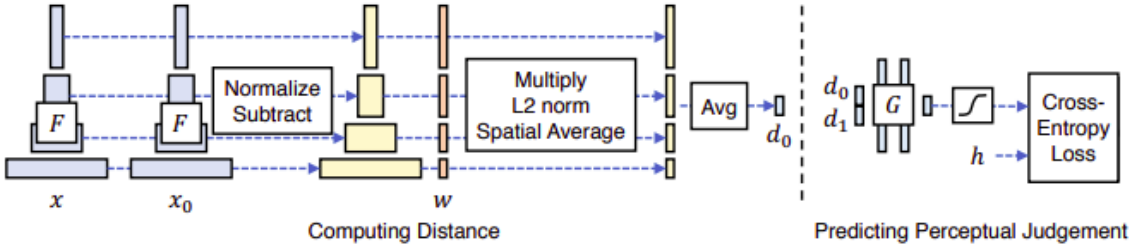
where $\alpha, \beta, \gamma > 0$ are parameters for weighting these components. When we set $\alpha = \beta = \gamma = 1$, we obtain

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{16}$$

Learned Perceptual Image Patch Similarity. Zhang et al. introduced a perceptual dataset containing 484k human judgements with parametrized distortions and real algorithm outputs [38]. They have shown that deep features trained with various levels of supervision (supervised, unsupervised, self-supervised) for different computer vision tasks work well as a perceptual metric. In other words, according to Zhang et al., features used for semantic tasks are also good at providing models of human perceptual behavior and outperform the previously used metrics such as PSNR and SSIM. Details of the metric is illustrated in Fig. 4.2.



(a) Preference of humans and different metrics for sample image patches



(b) Computing distance between two patches

Figure 4.2. LPIPS metric (taken from [38])

Perceptual Image Error Assesment through Pairwise Preference. According to Prashnani et al., comparing two given images and determining which one is more similar to a reference image is easier than assigning quality scores for each image [71]. Correspondingly, they introduced a dataset where the label of an image is the probability of being preferred over the other by humans. They employed a pairwise-learning framework to predict whether the distorted image will be preferred over the other. Hence, they have proposed a perceptual image-error metric that outperforms existing metrics with the observation that pairwise preference is a robust technique to train an error-estimation function.

Perceptual Index. According to Blau et al., there is a tradeoff between accuracy of reconstruction and perceptual quality [73]. Thus, they have used a perception-distortion plane for evaluating the super-resolution methods for the Perceptual Image Restoration and Manipulation (PIRM) challenge. They defined the perceptual quality as the visual quality of the reconstructed image regardless of its similarity to the ground truth. Thus, for the perception axis, they used perceptual index which is a combination of two no-reference image quality

metrics of Ma [39] and NIQE [77], defined as:

$$\mathrm{PI} = \frac{1}{2}\left((10 - \mathrm{MA}) + \mathrm{NIQE}\right) \tag{17}$$

NIQE metric proposed by Mittal et al. is based on a multivariate Gaussian model (MVG) fit of the natural scene statistics (NSS) [77]. They fit an MVG to the features extracted from a collection of natural images. Then, they fit another MVG to the NSS features extracted from a given image. These features are obtained from local image patches which are selected with respect to their sharpness. The quality of the image is measured by computing the distance between the two MVG models. Ma et al. suggested three types of properties as features [39] as shown in Fig. 4.3. First, they use the statistics of discrete cosine transform (DCT) coefficients to measure high-frequency artifacts. Second, they employ Gaussian scale mixture (GSM) model to extract global statistics information of the image. Third, principal component analysis (PCA) is applied to images and singular values are used to describe spatial discontinuities. Then, they regress these features to perceptual scores collected from human subject studies to obtain a no-reference metric.

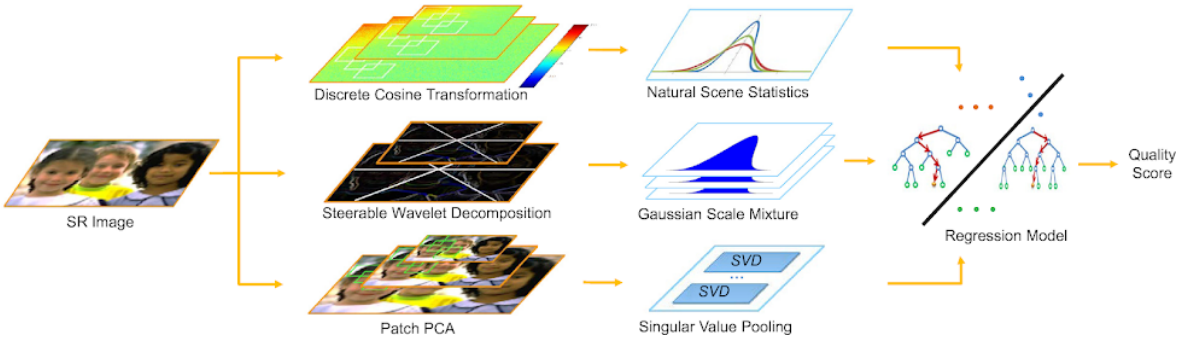
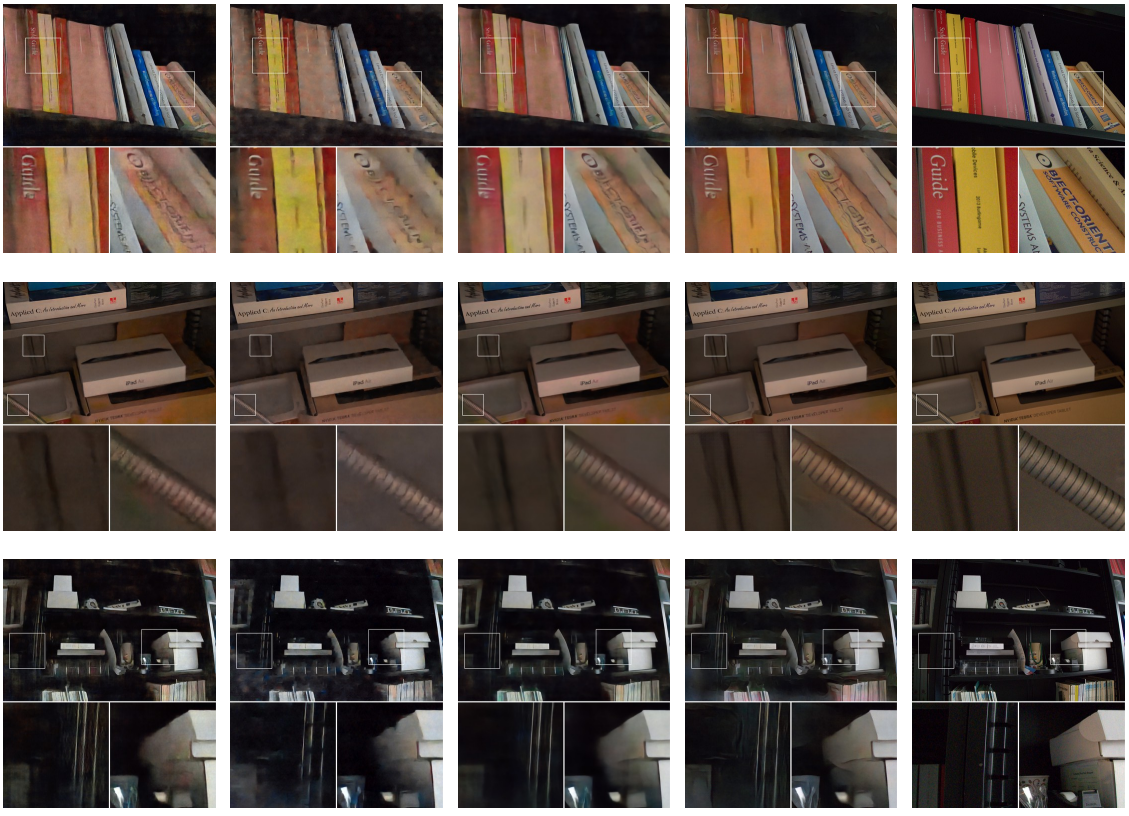


Figure 4.3. Ma metric (taken from [39])

4.4. Experimental Results

We first analyze the effectiveness of our coarse-to-fine strategy, and the performance gains achieved over the existing single image models. Fig. 4.4. shows visual comparison of our single image model against the state-of-the-art [28–30]. For the first image, the color of the books and the details of texts contained on the spines are better recovered by our model. For

the second image, the fine details are more visible and the edges are sharper (e.g. the lines on the wall and the cable) in our result. For the third image, our model greatly reduces the noise in the dark regions. Moreover, it is apparent that our approach preserves the edges better. Table 4.1. shows a quantitative analysis of our single image model on the SID dataset. Overall, our model outperforms the state-of-the-art in terms of PSNR and all perceptual metrics, LPIPS, PieAPP and PI, and gives competitive results in terms of SSIM. It should also be noted that our model achieves the highest PSNR on the dark images with $\times 250$ and $\times 300$ amplification ratios which are more challenging than the other subsets of $\times 100$.



(a) SID [28] (b) Maharjan et al. [29] (c) Zamir et al. [30] (d) Ours *(S)* (e) Ground truth

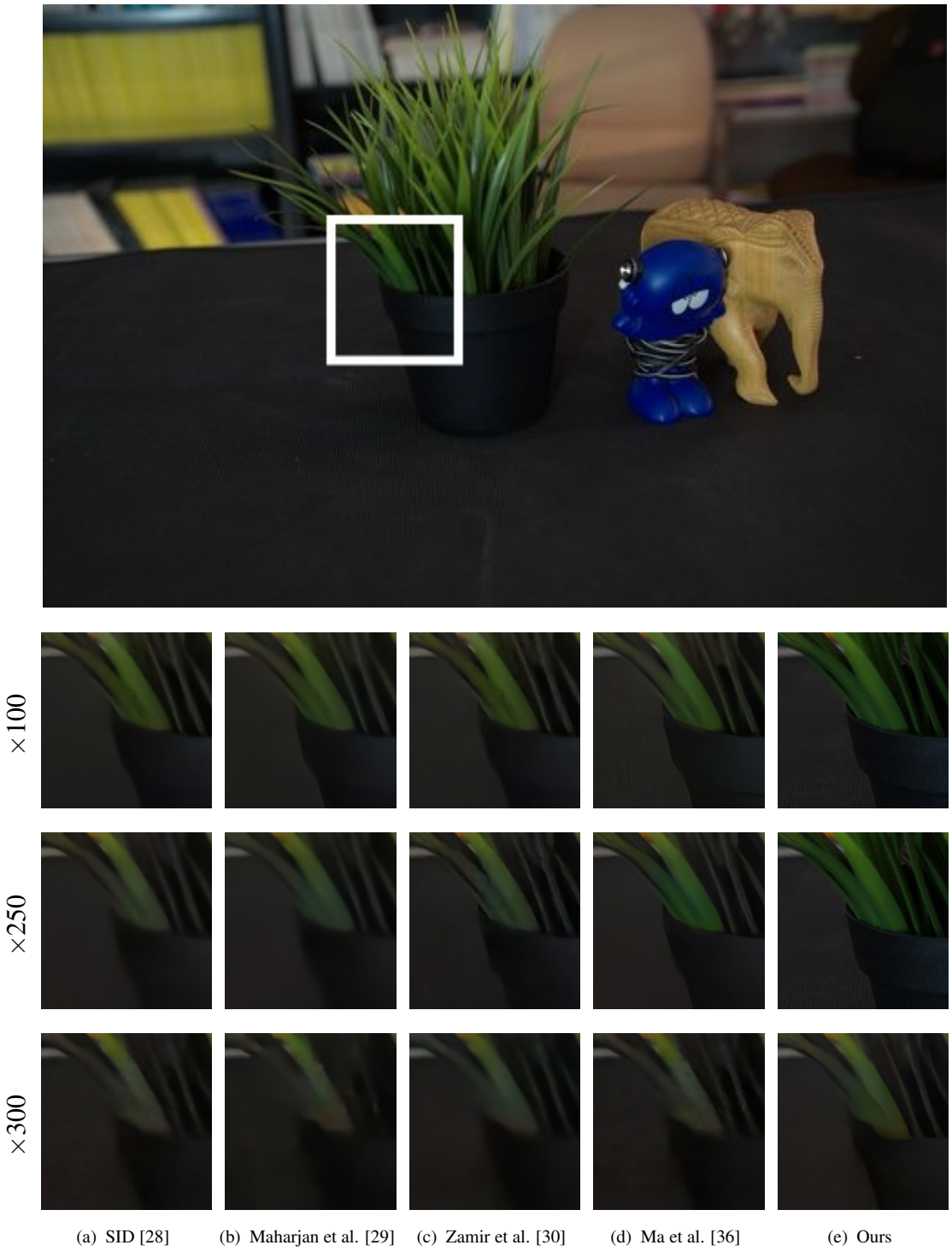
Figure 4.4. Qualitative comparison of our coarse-to-fine single image *(S)* method for enhancing extremely low-light images, compared against the state-of-the-art models that also process single image. From top to the bottom row, the amplification ratios are $\times 250$, $\times 100$ and $\times 250$, respectively.

Fig. 4.5. presents some visual results of our burst model, along with a performance comparison to the burst method of [36] and the ensemble versions of the single image methods [28–30]. As evident from the zoomed-in regions, our permutation-invariant CNN model

(a) SID *(E)* [28] (b) Maharjan et al. *(E)* [29] (c) Zamir et al. *(E)* [30] (d) Ma et al. *(B)* [36] (e) Ours *(B)* (f) Ground truth

Figure 4.5. Qualitative comparison of our burst *(B)* model for enhancing extremely low-light images, compared against the burst model by Ma et al. [36] and the ensemble versions *(E)* of the single image state-of-the-art models. From top to the bottom row, the amplification ratios are $\times 100$, $\times 300$ and $\times 300$ respectively.

can produce enhancement results with much sharper and well restored texture details. On the other hand, the ensemble methods all suffer from over-smoothing of the fine-scale details such as the thin lines on the mat and the printed characters on the spine of the book, and the textured regions like the green bush. The burst method of [36] does relatively better but its outputs are of low contrast. Moreover, we provide a comparison of our method with other methods for different amplification ratios in Fig. 4.6. As can be seen from the cropped region, our method better handles the color inaccuracies and the edges for higher amplification ratios, i.e, when the input images get darker. Table 4.2. clearly demonstrates the benefit of



(a) SID [28] (b) Maharjan et al. [29] (c) Zamir et al. [30] (d) Ma et al. [36] (e) Ours

Figure 4.6. Comparison of our method with other methods for different amplification ratios. When the input images are not very dark (e.g. with $\times 100$ amplification ratio), all of the methods give similar outputs. When the input images are darker and noisier (e.g. with $\times 300$ amplification ratio) our method better handles the color accuracy and the edges.

Table 4.1. Performance comparison of single image models on the SID dataset for different amplification ratios, with the best performing model highlighted with a bold typeface.

Ratio	Method	PSNR↑	SSIM↑	LPIPS↓	PieAPP↓	PI↓
×100	SID [28]	30.087	0.904	0.450	1.427	4.320
	Maharjan et al. [29]	**30.535**	**0.906**	0.448	1.250	4.481
	Zamir et al. [30]	29.922	0.895	0.465	1.310	4.518
	Ours *(S)*	30.464	0.905	**0.292**	**0.968**	**4.309**
×250	SID [28]	28.428	0.887	0.482	1.601	4.577
	Maharjan et al. [29]	28.787	**0.888**	0.488	1.443	4.961
	Zamir et al. [30]	28.254	0.878	0.462	1.462	4.956
	Ours *(S)*	**28.900**	0.884	**0.326**	**1.113**	**4.551**
×300	SID [28]	28.528	**0.870**	0.507	1.644	4.107
	Maharjan et al. [29]	28.382	0.868	0.516	1.645	4.523
	Zamir et al. [30]	28.441	0.860	0.494	1.520	4.479
	Ours *(S)*	**28.669**	0.863	**0.356**	**1.048**	**4.039**
All	SID [28]	28.976	**0.886**	0.482	1.564	4.319
	Maharjan et al. [29]	29.167	0.886	0.487	1.462	4.646
	Zamir et al. [30]	28.838	0.876	0.465	1.437	4.639
	Ours *(S)*	**29.290**	0.882	**0.327**	**1.087**	**4.281**

our approach that it performs the best in terms of all perceptual metrics, LPIPS, PieAPP, PI, and the PSNR metric.

In Table 4.3., we report the runtime performances of our single image and burst models in comparison with other competing methods. In particular, we measure the time taken to process a single image and also a burst of 4 images. Our experiments are conducted on a machine with an NVIDIA GeForce GTX 1080 Ti 11GB graphics card using 4256×2848 pixels images. For single image enhancement, our single image model is a bit slower than SID [28] and Zamir et al [30] due to its multi-scale architecture, though it gives better enhancement results as discussed before. For burst enhancement, our model achieves the best runtime performance, with 1.509 sec for a burst size of 4. This clearly demonstrates the advantage of having a shared decoder to process burst features, contrary to the competing approaches. We additionally report the runtime of our burst model to enhance a burst of 8 frames. As can be seen, the increase in the runtime is not linear in the number of processed images. We only

Table 4.2. Performance comparison of burst *(B)* and ensemble *(E)* models on the SID dataset for different amplification ratios, with the best performing model highlighted with a bold typeface.

Ratio	Method	PSNR↑	SSIM↑	LPIPS↓	PieAPP↓	PI↓
×100	SID *(E)* [28]	30.361	0.908	0.447	1.441	4.686
	Maharjan et al. *(E)* [29]	30.833	**0.909**	0.445	1.324	4.863
	Zamir et al. *(E)* [30]	30.120	0.898	0.430	1.335	4.776
	Ma et al. *(B)* [36]	30.429	0.908	0.423	1.312	4.295
	Ours *(B)*	**30.849**	**0.909**	**0.280**	**0.945**	**4.233**
×250	SID *(E)* [28]	28.915	0.893	0.480	1.622	5.313
	Maharjan et al. *(E)* [29]	29.289	0.893	0.480	1.525	5.609
	Zamir et al. *(E)* [30]	28.630	0.882	0.454	1.495	5.406
	Ma et al. *(B)*[36]	29.053	**0.896**	0.470	1.517	4.429
	Ours *(B)*	**29.479**	0.892	**0.313**	**1.063**	**4.366**
×300	SID *(E)* [28]	28.979	0.878	0.516	1.699	4.606
	Maharjan et al. *(E)* [29]	28.783	0.875	0.520	1.744	5.003
	Zamir et al. *(E)* [30]	28.750	0.866	0.500	1.581	4.805
	Ma et al. *(B)*[36]	29.078	**0.884**	0.467	1.464	4.018
	Ours *(B)*	**29.232**	0.877	**0.322**	**1.048**	**3.923**
All	SID *(E)* [28]	29.383	0.892	0.484	1.596	4.850
	Maharjan et al. *(E)* [29]	29.568	0.891	0.485	1.548	5.148
	Zamir et al. *(E)* [30]	29.132	0.881	0.462	1.480	4.983
	Ma et al. *(B)*[36]	29.485	**0.895**	0.455	1.433	4.232
	Ours *(B)*	**29.804**	0.891	**0.306**	**1.021**	**4.157**

Table 4.3. Runtime analysis for single image and ensemble/burst models. The fastest model is indicated with a bold typeface. Running times are in seconds.

Method	1 frame	4 frames	8 frames
SID [28]	**0.424**	1.648	–
Maharjan et al. [29]	2.287	3.045	–
Zamir et al. [30]	**0.424**	1.648	–
Ma et al. [36]	–	2.001	–
Ours *(S)*	0.597	1.889	–
Ours *(B)*	0.597	**1.509**	2.413

observe $1.6\times$ increase when the burst size is doubled from 4 to 8. It should be noted that for the case of the burst size of 8, we were unable to report runtimes of the competing models here as enhancing these frames within a single batch by these models exceed the limits of our GPU memory.

Our model is entirely trained on the Sony dataset of SID [28] containing images captured by the Sony α7S II sensor. To demonstrate that our learned models can (partly) generalize to other camera sensors, in Fig. 4.7. and Fig. 4.8., we present example outputs of our single and burst image models on extremely dark photos taken with cameras of an iPhone 6s and an iPhone SE, respectively. Once again, Fig. 4.7. demonstrates that our model reduces the noise better than the state-of-the-art models [28–30], while accurately improving the texture details of the flower and the leaves. Similarly, Fig. 4.8. shows the cross-sensor generalization capability of our burst model. Our method clearly produces a better result than both the traditional camera pipeline[2] and SID [28] in that it recovers the details of the water hose and the leaves of the tree more accurately.

[2] `https://github.com/letmaik/rawpy`

SID [28]

Maharjan et al. [29]

Zamir et al. [30]

Ours (single)

Figure 4.7. Enhancement results of a raw image captured by an iPhone 6s using 1/20 sec exposure time and 400 ISO. Our proposed single image enhancement model provides better noise reduction with more structural details, in comparison to the prior approaches.

(a) Traditional Pipeline *(Ensemble)* (b) SID *(Ensemble)* [28] (c) Ours *(Burst)*

Figure 4.8. Enhancement results on a burst of 8 raw images taken with an iPhone SE with 1/10 sec exposure time and 400 ISO. Resulting images obtained by (a) averaging over the traditional pipeline, (b) averaging over the SID [28] predictions, (c) our burst model.

4.5. Ablation Study

To evaluate the effectiveness of our approach in more detail and to better understand the effects of the loss functions and also the contribution of the burst size to the overall quality, we conducted an extensive series of ablation tests.

Losses. As mentioned before, the loss function we used to train our networks consists of two complementary loss terms. The first term is the pixel-wise L_1 loss which is used to improve the accuracy of reconstructing a long-exposure image. The second term, on the other hand, is comprised of the contextual L_{CX} loss function, which is utilized to improve the perceived quality of the end result.

In Table 4.4., we quantify the effect of using the contextual loss, as opposed to the perceptual loss, in conjunction with the pixel-wise L_1 loss. First of all, the burst model trained with only L_1 loss results in higher PSNR and SSIM but relatively lower perceptual quality, which is in line with the previous observations [38, 73]. Conversely, using only L_p or L_{CX} loss gives higher perceptual but lower PSNR/SSIM scores. In that sense, adding either L_{P} or L_{CX} to our objective function provides a good tradeoff between pixel-wise and perceptual metrics. To inspect which one is better, we also qualitatively analyze the contribution of incorporating the perceptual loss L_{P} or the contextual loss L_{CX}. As demonstrated in Fig. 4.9., either L_{P} or L_{CX} allows improving the perceived quality of the end-result. The resulting images have more realistic fine-scale details and texture while avoiding over-smoothing. To our interest, however, the network trained with the contextual loss tends to better recover the thin structures, especially at the darker regions, as compared to the others. In Fig. 4.10., we provide some additional results from training with L_1, L_p or L_{CX} losses separately. For the first image, contextual loss produces a better result in terms of texture and color bleeding. Meanwhile, it is also observed that contextual loss tends to introduce imaginary texture in some flat regions where L_1 and perceptual losses smooth the corresponding regions as illustrated in the second image.

Table 4.4. Effect of the loss functions on the performance of the proposed burst enhancement model.

Method	PSNR↑	SSIM↑	LPIPS↓	PieAPP↓	PI↓
L_1	29.843	**0.898**	0.417	1.364	4.252
L_p	25.347	0.668	0.300	1.340	4.834
L_{CX}	20.226	0.453	0.343	1.213	4.563
$L_1 + L_{\text{P}}$	**29.895**	0.894	**0.274**	1.053	4.593
$L_1 + L_{\text{CX}}$	29.804	0.891	0.306	**1.021**	**4.157**

(a) L_1 (b) $L_1 + L_{\text{P}}$ (c) $L_1 + L_{\text{CX}}$

Figure 4.9. Enhancement results of our method with different loss functions. Utilizing the combination of contextual loss and pixel-wise loss gives visually more pleasing results, as compared to using the pixel-wise loss together with and without the perceptual loss.

Burst Processing. In Fig. 4.11., we analyze how the number of frames in the burst sequence affects the performance of our model. Here, we provide the results obtained with a single input image and the burst sizes of four and eight frames. As can be seen from the zoomed-in results, the output quality improves with an increasing number of the burst images – the method gets much better at preserving texture details and thin structures. In Fig. 4.12., we also compare our (set-based) burst method with the ensemble of our single image model (i.e.,

(a) L_1 (b) L_P (c) L_{CX}

Figure 4.10. Enhancement results of our method with individual loss functions. While perceptual and L_1 losses are more accurate in flat regions, the contextual loss is better at dealing with thin structures and color bleeding.

Table 4.5. A quantitative comparison of the proposed burst model with the ensemble of the single image model for varying number of burst images.

Method	PSNR↑	SSIM↑	LPIPS↓	PieAPP↓	PI↓
Ours *(S)*	29.290	0.882	0.327	1.087	4.281
Ours *(E)* (4 frames)	29.706	0.888	0.329	1.121	4.762
Ours *(E)* (8 frames)	29.738	0.889	0.332	1.126	4.716
Ours *(B)* (4 frames)	29.742	0.890	0.313	1.034	4.197
Ours *(B)* (8 frames)	**29.804**	**0.891**	**0.306**	**1.021**	**4.157**

processing each image in the burst separately and then taking the average of individual outputs). Fusing burst images at the feature level is evidently much more effective. Additionally, in Table 4.5., we quantitatively evaluate the performance of these alternative strategies[3]. Our burst model gets better scores across all metrics as compared to the ensemble approach, even when using only half of the burst images.

[3]As mentioned before, the burst sizes for the images in the Sony dataset vary between 2 and 10. Here, we report the results obtained using at most four or at most eight burst frames.

(a) Single image (b) 4 frames (c) 8 frames (d) Ground truth

Figure 4.11. Effect of the burst size. As can be seen, as we increase the number of images in the burst sequence, the enhancement quality of our burst model improves further.

4.6. Limitations

Our approach does have a few limitations. First and foremost, our burst approach might struggle with the burst sequences having large motion changes or camera-shake since it is trained on a dataset where the burst frames are spatially aligned. We present such an example in Fig. 4.13., in which our burst model introduces some unintuitive edges and blurry textures because of the misalignment while the single image model produces much sharper output. Second, as illustrated in Fig. 4.14., our model may sometimes hallucinate non-existing high-frequency details. We suspect that this is caused by the excessive noise in the raw images

(a) Ours (ensemble) (b) Ours (burst)

Figure 4.12. A comparison between our burst model and the ensemble version of our single image model for a burst size of 8 images. Our set-based approach, which performs fusion at the feature-level, gives perceptually better enhancement results.

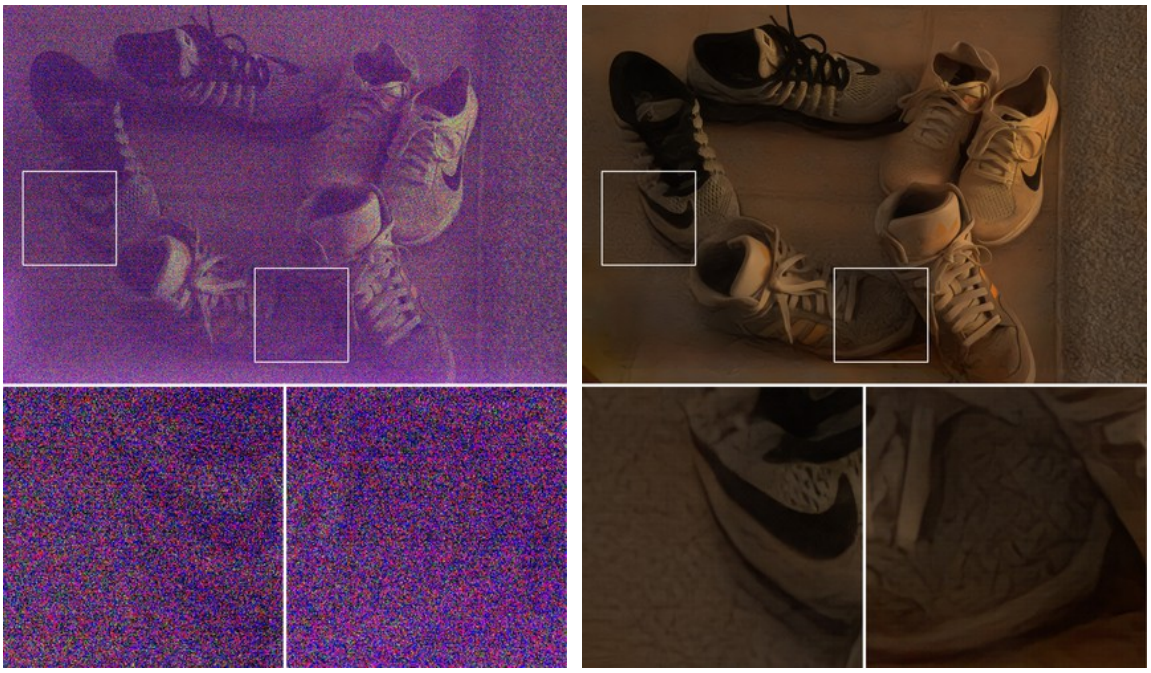
and may be alleviated to some extent by better modeling of the sensor noise. Third, our framework does not explicitly learn to perform white balance correction and tone mapping, and this somewhat affects the results. In an attempt to address this, we employ an additional post-processing step. In particular, we first apply the white balance correction method proposed in [78] to our result. Then, we adjust highlights and shadows using the Core Image API by Apple[4]. Finally, we merge this image with the white-balanced image by using the exposure fusion method by Mertens et al. [79] to obtain a tone-mapped image. Fig. 4.15. presents the result of this post-processing step on a sample dark input image. It is evident that this post-processing strategy leads to a visually more pleasing image with vivid colors, further improving the perceived quality of the enhanced image.

[4]Documentation of the API can be found at `https://developer.apple.com/documentation/coreimage`.

(a) Ours (single)

(b) Ours (burst)

Figure 4.13. A limitation of the proposed burst model. Our model might generate unintuitive edges and blurry textures when the burst frames are not spatially well-aligned.



(a) Traditional pipeline (b) Ours (burst)

Figure 4.14. Another limitation of the proposed approach. Our model may sometimes hallucinate false high-frequency details for extremely noisy regions.

(a) Ours (b) Ours + Post-process

Figure 4.15. Effect of the post-processing procedure applied to the result of our model for a low-light image captured with 0.1 sec exposure. Post-processing further improves the perceived quality of the enhanced image.

5. CONCLUSION

In this thesis, we tackle the problem of learning to generate long-exposure images from a set of extremely low-light burst images. Starting from the measurement of how dark a scene is, we first analyzed how capturing images in the dark affects the brightness of the resulting image. It is clear that current photography techniques such as lengthening the exposure time or using the camera flash do not solve the problems in the low-light images. For instance, long exposure images are not practical and may result in blurry images because of the movement of the camera or scene elements. Using a camera flash as an external light source may increase the brightness but the brightness of the scene elements becomes inversely proportional to their distance to the camera. Thus, there have been various studies trying to reduce the image noise and improve the final image quality for the low-light images. The recent trend in both image denoising and enhancement is to develop a learning-based model for the real-life, raw images that contain real image noise. Besides, learning-based methods have also shown promising results for the burst image denoising and deblurring problems.

Motivated by the recent advances, we developed a new deep learning based approach that incorporates a coarse-to-fine strategy to better enhance the details of the output. Moreover, we extended this network architecture to work with a burst of images via a novel a permutation invariant CNN architecture, which efficiently processes the exchanged information between the features of the burst frames. Our experimental results demonstrate that our burst method achieves higher quality results than the existing state-of-the-art models. More specifically, it better captures finer details, texture and color information, and reduces noise to a greater extent.

As for future work, our results suggest that there is still much room for improvement, especially for dynamic scenes. In that sense, an interesting future research direction is to extend the proposed framework to videos with moving objects or fast camera motions where capturing temporal relationships between succeeding frames is crucial.

REFERENCES

[1] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T Barron, Dillon Sharlet, Ryan Geiss, et al. Handheld mobile photography in very low light. *ACM Trans. Graphics*, **2019**.

[2] Samuel W Hasinoff. Photon, poisson noise. *Computer vision: a reference Guide*, **2014**.

[3] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *CVPR*. **2019**.

[4] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Trans. Graphics*, **2016**.

[5] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM Trans. Graphics*, **2004**.

[6] Daisuke Sugimura, Takuya Mikami, Hiroki Yamashita, and Takayuki Hamamoto. Enhancing color images of extremely low light scenes based on rgb/nir images acquisition with different exposure times. *IEEE Trans. Image Process.*, **2015**.

[7] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *CVPR*. **2005**.

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans. Image Process.*, **2007**.

[9] Hossein Talebi and Peyman Milanfar. Global image denoising. *IEEE Trans. Image Process.*, **2013**.

[10] S Grace Chang, Bin Yu, and Martin Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. Image Process.*, **2000**.

[11] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, **2006**.

[12] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, **1992**.

[13] V. Jain and S. Seung. Natural image denoising with convolutional networks. In *NeurIPS*. **2009**.

[14] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *NeurIPS*. **2012**.

[15] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.*, **2017**.

[16] Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Trans. Image Process.*, **2018**.

[17] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *ICML*. **2018**.

[18] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *CVPR*. **2019**.

[19] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. In *NeurIPS*. **2019**.

[20] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *CVPR*. **2019**.

[21] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, **2017**.

[22] Li Tao, Chuang Zhu, Guoqing Xiang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Llcnn: A convolutional neural network for low-light image enhancement. In *VCIP*. **2017**.

[23] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. MBLLEN: Low-light image/video enhancement using CNNs. In *BMVC*. **2018**.

[24] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*. **2019**.

[25] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*. **2018**.

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

[27] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. *arXiv: 2001.06826*, **2020**.

[28] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*. **2018**.

[29] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via residual learning. In *ICME*. **2019**.

[30] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Fahad Shahbaz Khan, and Ling Shao. Learning digital camera pipeline for extreme low-light imaging. *arXiv: 1904.05939*, **2019**.

[31] Toni Buades, Yifei Lou, Jean-Michel Morel, and Zhongwei Tang. A note on multi-image denoising. In *International Workshop on Local and Non-Local Approximation in Image Processing*. **2009**.

[32] Neel Joshi and Michael F Cohen. Seeing mt. rainier: Lucky imaging for multi-image denoising, sharpening, and haze removal. In *ICCP*. **2010**.

[33] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun. Fast burst images denoising. *ACM Trans. Graphics*, **2014**.

[34] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *CVPR*. **2018**.

[35] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising. In *ECCV*. **2018**.

[36] Lan Ma, Di Zhao, Songnan Li, and Dahai Yu. End-to-end denoising of dark burst images using recurrent fully convolutional networks. In *VISIGRAPP*. **2020**.

[37] Marc Levoy. Night sight: Seeing in the dark on pixel phone. *Google AI Blog*, **2018**.

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*. **2018**.

[39] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, **2017**.

[40] Shuhang Gu and Radu Timofte. A brief review of image denoising algorithms and beyond. In *Inpainting and Denoising Challenges*. **2019**.

[41] P. Chatterjee and P. Milanfar. Is denoising dead? *IEEE Trans. Image Process.*, **2010**.

[42] Robert Hummel. Image enhancement by histogram transformation. *Computer Graphics and Image Processing*, **1977**.

[43] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics Gems IV*. **1994**.

[44] Haidi Ibrahim and Nicholas Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Trans. Consum. Electron.*, **2007**.

[45] Tarik Arici, Salih Dikbas, and Yucel Altunbasak. A histogram modification framework and its application for image contrast enhancement. *IEEE Trans. Image Process.*, **2009**.

[46] Edwin H. Land. The retinex theory of color vision. *Scientific American*, **1977**.

[47] Michael K Ng and Wei Wang. A total variation model for retinex. *SIAM J. Imag. Sci.*, **2011**.

[48] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*. **2016**.

[49] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.*, **2017**.

[50] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.*, **1997**.

[51] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. **2015**.

[52] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *NeurIPS*. **2016**.

[53] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*. **2016**.

[54] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *CVPR*. **2020**.

[55] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *CVPR*. **2020**.

[56] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Maintaining natural image statistics with the contextual loss. In *ACCV*. **2018**.

[57] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *ICCV*. **2019**.

[58] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *ICCV*. **2019**.

[59] Mauricio Delbracio and Guillermo Sapiro. Hand-held video deblurring via efficient fourier aggregation. *IEEE Trans. Comput. Imag.*, **2015**.

[60] Miika Aittala and Frédo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *ECCV*. **2018**.

[61] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*. **2017**.

[62] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*. **2018**.

[63] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. In *ICML*. **2017**.

[64] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*. **2016**.

[65] Robert Gens and Pedro M Domingos. Deep symmetry networks. In *NeurIPS*. **2014**.

[66] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *NeurIPS*. **2017**.

[67] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *ECCV*. **2018**.

[68] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, **2016**.

[69] Tommi Kärkkäinen, Karl Kunisch, and Kirsi Majava. Denoising of smooth images using l 1-fitting. *Computing*, **2005**.

[70] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *Journal of Vision*, **2016**.

[71] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *CVPR*. **2018**.

[72] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*. **2015**.

[73] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCV*. **2018**.

[74] Rafael C Gonzalez and Richard E Woods. *Digital image processing*. Pearson Education, **2018**.

[75] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Process Mag.*, **2009**.

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

[77] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, **2012**.

[78] Mahmoud Afifi, Brian Price, Scott Cohen, and Michael S Brown. When color constancy goes wrong: Correcting improperly white-balanced images. In *CVPR*. **2019**.

[79] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer Graphics Forum*. **2009**.