Improving Mobile Phone Image Quality by Deep Learning

Cen Huang
The Chinese University of Hong Kong
hc015@ie.cuhk.edu.hk

Yicun Liu
The Chinese University of Hong Kong
ly116@ie.cuhk.edu.hk

Abstract

Despite the rapid progress in the processing power of smart phones, the improvement of mobile device's image quality has long been slow. Due to the strait physical limitations in sensor size and aperture optical structure, it remains challenging for mobile devices to compete against those DSLR cameras. In this thesis, we bridge the gap by introducing our deep neural enhancement framework which generally learns the enhance transformation from mobile-phone-quality images to DSLR-quality images. To solve the imperfect alignment problem of our training data, we proposed two methods: the first method goes through a two-phase enhancement, on the one hand, it gradually refines the detailed texture by implementing DPED end-to-end frameworks, one the other hand, it adopts style-transfer techniques and implement the locality semantic enhancement by feed-forward structures. The second method first solves a intrinsic decomposition and then enhances the shading, reflectance, and lighting components respectively. This framework is trained with three datasets with different concentration in in color, texture and noise, and focuses only on its meaningful components in the intrinsic decomposition. To that end, we establish a framework which first learns intrinsic image decomposition on self-supervised manner, then apply a multi-domain image translation architecture to learn a comprehensive enhance transformation from those partially enhanced datasets. [1]

1. Introduction

Improving quality of images produced by mobile devices, which can be regarded as many sub-problems such as image super-resolution, image denoising, image deblurring, contrast enhancement, chromatic balancing, has been an emerging researching and industrial area in recent years. Due to the size limitation of mobile optics, together with the hand shaking from the photographer, images shot by mobile users are mostly undesirable in structural details and textural information. Moreover, images taken in the low-light environments by mobile devices are usually lack of details and within low-dynamic range, leading to the unsatisfying performance in the overall visual assessment.

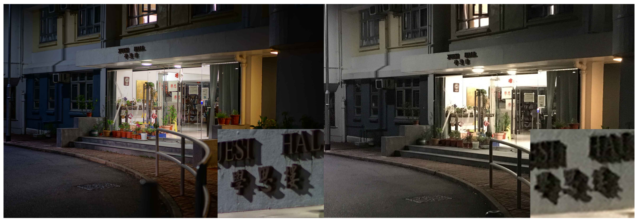


Figure 1. Quality comparison of DSLR Image(left) and Mobile Phone Image(right): It can be observed that DSLR image has more texture and structure details than the image shot by iPhone. Also, for high-contrasting scene, our sampling DSLR equipped with HDR has better dynamic range, showing especially better quality and more detail in the shadow area of the image.

So far, there exist a number of precedent methods specialized for one of those sub-problems, but still, no universal solution conjugating all those sub-domain solutions has been discussed. Contrast focused approaches as histogram equalization can only adjust the global contrast and maximize the information entropy, which brings about limited improvement in detail reconstruction and hue balancing. Chroma focused methods aim to sample and adjust the hue distribution of the images, targeting at remapping the color space of the original pictures into a more attractive one. Those methods are usually specially designed for specific scenarios of images, with considerable computational cost and usability constraints.

Until very recently, Convolutional Neural Networks started to demonstrate its superiority in image enhancement task. Different from previous methods with a specific focus, CNN based methods mostly benefit from using a large volume of images in the training process, to learn the corresponding high-quality features. Up to now, the most impressive sub-task benefited from CNN is super-resolution proposed by Dong at [7]. Although quantitative measurements like PSNR and SSIM indicate the significant improvement in recent works, most tests are still based on general downgrading algorithms such as bicubic downsampling

[1] The two authors are of equal contribution.

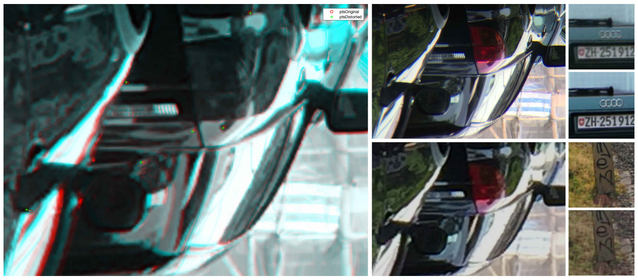


Figure 2. Difficulties in image registration due to different sensors and optics: The left image is the result of registration, in which the red and blue area indicate the imperfect alignment. It can be observed that even with camera registration, misalignment still exists, this problem is also reflected in another similar dataset[2].

and Gaussian blurring to synthesize low-quality samples in the training step. Those techniques are not suitable for real-world image enhancement tasks: visual quality defined in real-world images is far more than maintaining sharpness in edges, and most importantly, it is infeasible to decompose the low-quality factors and treat them respectively from a low-quality image.

When it comes to the real-world image enhancement task, the most intuitive approach is sampling high quality and low-quality images separately from DSLRs and mobile phones, then inputting them into an end-to-end training procedure to learn the contributing features of high visual quality. The motivation for making this initial attempt is straightforward: given two images of the same scene, with high quality and low quality respectively, learning the enhancement transformation which explicitly improves those sub-domain qualities is plausible. However, there are two fundamental challenges in practice: paired image misalignment and unpaired training ambiguity.

Misalignment in Sampled Images - In real-world cases, images taken by mobile phones often suffer from different extents of distortion and dispersion, preventing accurate alignment without appropriate undistortion in the prior processing step. Even undistortion itself is challenging: The distortion parameter matrix of a certain camera is not singular, considering the movement of relative position in the lens when focusing at different ranges, no unique solutions can be constructed from such a rank-deficient matrix. Even if the corrected images appear natural in global views, eventual registration by existing MATLAB algorithms fails to provide a reasonable pixel-wise mapping. Consequently, the pixel-wise loss function, which is indispensable to low-level vision tasks, can no longer contribute to our training process.

Lack of Semantic Supervision in Pairwise Training - When applying the end-to-end pairwise training strategies with simple $l1$ or $l2$ regularization, an essential problem is worth consideration: How to divide high-quality and low-quality image pairs into semantic-meaningful patches for network training? In most cases, for the sake of simplicity, image patches are randomly sampled and used for training, preventing accurate semantic information from being transferred into the neural networks, which causes the loss of details in generated results. Even strange artifacts can happen in paired learning: The structural information of one object can be copied into multiple patches and get enhanced differently, and the border between different patches can be concatenated unnaturally. Consequently, even if the partial details are greatly improved, the global image quality may still not be visually pleasing.

Ambiguity and Divergence in Unpaired Training - Without the implementation of a pixel-wise loss function, it appears exceptionally difficult to progress the training with tight constraints. However, there also exists weakly supervised approaches in using Generative Adversarial Network [45, 20] and Style Transfer methods [14]. However, this kind of methods requires highly selective training strategies and carefully-designed network structures Due to the weakly supervised manner, the convergence direction of network training will probably be ambiguous and underdetermined. Also, especially for night-view images, weak constraints would always give rise to clear noise patterns in the generated images.

In this thesis, we particularly focus on solving the three aforementioned problems when using real-world images to learn the enhancement transformation. To deal with misalignment in paired training strategy, we propose the pixel-shift insensitive color loss to alleviate the situation. To compensate for the semantic loss, we propose the sentiment-based locality enhancement to reinforce semantic details of individual regions. Coping with the ambiguity of training directions, we combine the paired DPED method with the weakly supervised structures, which establishes a two-phase enhancement procedure.

The main contribution of our work can be summarized in the following perspectives:

- First, we thoroughly investigate previous algorithms for each sub-task of the overall image enhancement, such as super-resolution, deblurring, denoising and style transfer. Based on the performance in those sub-areas, we assess the dominating factors for image quality and gain further insight when designing our universal framework concerning all those sub-problems. p

- Second, we proposed a novel method which divides the low-light image enhancement into two parts: detailed texture enhancement and global style enhancement. We first devised an enhanced DPED structure to impose the pixel-wise constraints, while intermediate generative images are then incorporated with the professional style images and enhanced in a weakly supervised manner. Different from the previous method which only accomplishes their aesthetic enhancement on one label, our method aims to explicitly achieve feature reinforcement and style augmentation under the assistance of both ground truth image and reference style image.

- Third, we explore the effect of combing image decomposition and multi-domain image to image translation. The first part is self-supervised intrinsic image decomposition, which learns from unlabeled real-world intrinsic data. With the shading, reflectance, and lighting components in hand, we use multi-domain GAN architecture to learn the partial enhancement of existing datasets which are specialized in color enhancement, texture enrichment, and noise reduction. We train this two parts together and aims at learning the full enhancement transformation for mobile image input.

- Fourth, we contribute a comprehensive low-light mobile-DSLR dataset. The dataset contains RAW file taken by ourselves and various JPEG file downloaded from FLICKER. All the RAW files we collected come with a human-retouched version from professional photographers. The dataset contains over 100 scenes of night view and includes over 1000 images.

2. Related Work

2.1. Image Super-Resolution

Image resolution has always been a critical criterion for image structure and edge enhancement. Pioneering approaches use interpolation techniques in sampling theory [27, 43] and adopt statistics of natural images to reconstruct realistic textures in the output images [44, 37]. Advanced works often targeted at learning the mapping function from I^{LR} to I^{HR} with techniques like sparse coding [42, 41] and neighborhood embedding [6].

Recently, the superior performance of deep neural networks also has significant impact on low-level vision tasks such as super-resolution. SR using convolutional neural network was first proposed by Dong at [7], with the accelerated version with deconvolution at [8]. Kim implemented residue learning at the task of SR [24] and designed more in-depth architecture [28] which dropped the batch normalization layer to achieve state-of-art performance in the most recent NTIRE super-resolution challenge [38]. Apart from Single Image Super-Resolution (SISR), other multi-frame solutions have also been proposed for video super-resolution [25, 26]. When adopting super-resolution on one specific frame, the multi-frame methods predicts the next framework based on prior knowledge, by which achieves high inter-frame consistency and structure fidelity in the resulted videos.

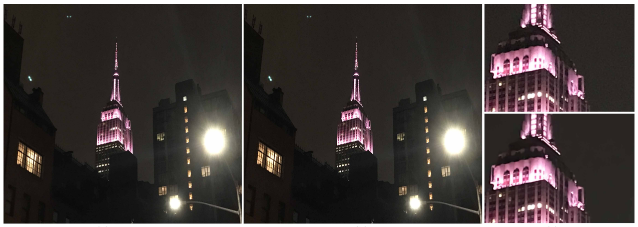


Figure 3. Side-by-side denoise comparison: (a) Denoised result using real-world data for training (training data generated by multi-frame denoising). (b) Denoised result using synthesis data for training (training data generated by adding Gaussian noise to clean images). (c) Close up comparison, where on the top is the closeup of (b), on the bottom is the closeup of (a). It can be observed that using real-world data for training leading to obvious improvement in the single sub-task of denoising, which could indicate that the potential problem of using synthesized training data.

Regardless the emerging progress in the subtask of super-resolution, until now, existing works of super-resolution take simple downsampling methods like the bicubic interpolation to synthesize the lack of details in real-world scenarios. Nevertheless, such deterministic downgrading methods could be problematic. In the side-by-side comparison of DSLR and mobile phone images, lacking in details might not correspond with what people expect: There exist salient differences between the images of mobile phone and downsampled version of DSLR, which indicates the drawback of generating training data by theoretical downsampling.

2.2. Image Denoising and Restoration

Image denoising and restoration is another critical issue in image enhancement task for images generated from tiny mobile sensors under low-light condition. In the past decades, extensive methods of denoising and image restoration methods have been looked into. Early methods like BM3D [11] and other dictionary-based methods [40, 9, 17] have demonstrated promising capabilities in image restoration tasks including denoising. Later, Burger at [4] challenged BM3D by using the Multi-Layer Perceptron to learn the end-to-end denoising transformation from noisy images to their clean versions. Stacked denoising auto-encoder then exploited the unsupervised pre-training to minimize the reconstruction error for better-denoised images [39]. Deep convolutional neural networks later improve the previous result by using deep auto-encoders with skip connections [31].

However, most of the previous denoising frameworks are not based on real-world cases. Instead of conducting experiments on low-light images with severe noise, nearly all of the learning methods simply utilize Gaussian white noise or Poisson noise to generated the noisy counterparts from the clean images. Although training on synthesized images achieves impressive results in quantitative measurement such as PSNR, but still performing poorly when directed adopting on the real-world noisy images generated by mobile phones.

2.3. Color and Contrast Enhancement

Image colorization and contrast enhancement are especially important for low-light image enhancement. Since in the night-view dataset, we always find that the images are in low dynamic range and always have intensity around 0-20. The most commonly used method is through Histogram Equalization(HE)[34]. Histogram Equalization is built on the assumption that when we convert the probability density function into a uniform distribution, it will maximize the information entropy and achieve the greatest contrast. HE remaps the gray levels of the image histogram based on the input cumulative distribution function. After the remapping, the image histogram will always fill in the whole dynamic range.

2.4. Image-to-Image Translation

The process of enhancing a low-quality image to the high-quality one can be considered as the specialized task of image-to-image translation. Before us, precedent work has introduced the idea of image-to-image translation in [18], which employed the non-parametric texture model [13] on an input-output image pair. Recent methods used end-to-end training to learn the translation function from the input-output pairs [29]. Pix2Pix was then proposed by Isola at [21], which employed the conditional generative adversarial network [15] to the learn the translation mapping between the pairs.

Recently, with GAN-based conditional image generation being actively studied, multi-domain image-to-image translation has become possible. Star-GAN has been newly proposed by [10], and is capable of learning the mappings among multiple domains, using a mask vector on a single generator and discriminator.

2.5. Photorelistic Style Transfer

Photographic style transfer techniques often seek to transfer the style of the original image to another input image, as if it is taken under different illumination, time of day or weather. Global style transfer process an image by applying a spatially-invariant enhance function to handle tone curves and global color shifts. [12] While the local style transfer focuses more on sentiment regional transfer such as time-of-day hallucination [22], weather-to-season change [22] and painterly stylization [1]. One possible real-world application would be deep photo style transfer [14], which focuses on the styling the original image while keeping the result photorealistic. One of our proposed methods will utilize the idea of deep style transfer and set the reference image as an ideal aesthetic reference while the original image as the intermediate result from phase-1 enhancement.

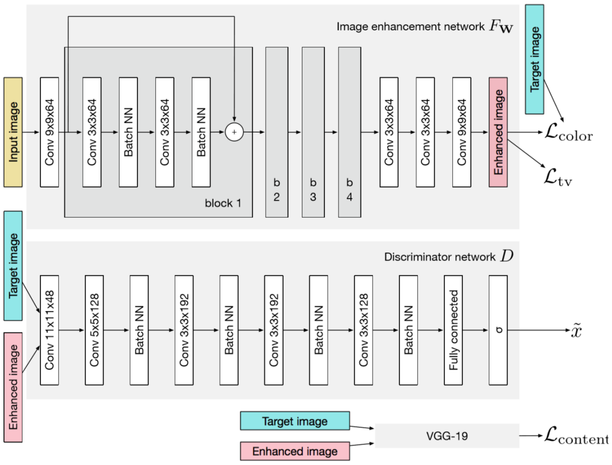


Figure 4. The overall architecture of proposed paired learning procedure: Training network consists of 1 9×9 layer followed by 4 residual blocks, then concatenated with 2 3×3 layers and 1 9×9 layer. Discriminator network consists of a similar structure of VGG networks and has a fully-connected layer at the bottom. VGG-19 network is used for calculating the content loss between generated image and its ground truth.

2.6. Intrinsic Image Decomposition

From the perspective of image formation, each image can be defined as the pixel-wise producty of the Albedo $\mathcal{A}$, which defines a surface's base color and shading $\mathcal{S}$, which defines the captured influence of light reflection and shadowing on each pixel: $\mathcal{I} = \mathcal{A} \cdot \mathcal{S}$. Extensive work was focused on how to recover $\mathcal{A}$ and $\mathcal{S}$ correctly given an color image input $\mathcal{I}$[16]. Consider the real-world ground truth for this topic is extremely scarce, there exists only alternative human-labeled intrinsic evaluation for intrinsic[3]. Until very recently, only computer-generated intrinsic dataset is available. The paper [23] proposed a method to learn the intrinsic without ground truth in real life, using two VAEs with cycle loss, operated in a self-supervised manner.

Although the intrinsic of images indeed reveals physical attributes like reflectivity, shading in different components, there is no previous work utilizing intrinsic decomposition in image enhancement problem. We describe the image enhancement problem as a 'decompose-enhance-reconstruct' process and utilize the self-supervised intrinsic model to inference the real-life intrinsic for the mobile image.

3. Methods

We propose two methods to enhance the mobile image and learn the transformation. The first method is Multi-Phase Semantic Enhancement (MPSE) and the second method is Multi-Domain Intrinsic Enhancement (MDIE).

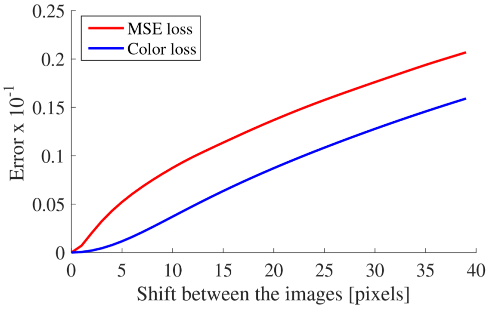


Figure 5. Comparison between responses to small image pixel-shift of MSE loss and Color loss. After applying the Guassian filter in the original feature maps, the color loss is more pixel-invariant to the MSE loss

3.1. Multi-Phase Semantic Enhancement

Our first algorithm focuses on image enhancement in multiple phases: paired pixel-wise learning, unpaired semantic style reinforcement, and post-processing optimization. We seek to improve the image quality from the microscopic view(pixels, textures) to macroscopic view(semantic regions, contrast, image style).

3.1.1 Paired Pixel-wise Learning

Color Loss Consideration As aforementioned in the misalignment problem in the introduction, paired images are taken from mobile devices and DSLR professional cameras, though shot on the same scene, are impossible to achieve perfect alignment. The fundamental strategy is to abandon the per-pixel loss and to scan the original feature maps by Gaussian filters with pre-computed Gaussian weights, resulting in invariant blurred color features. More specifically, input image X and Y will be converted to blurred image X_b and Y_b.

$$X_b(i,j) = \sum_{m,n} X_b(i+m, j+n) \cdot G_(m,n)$$

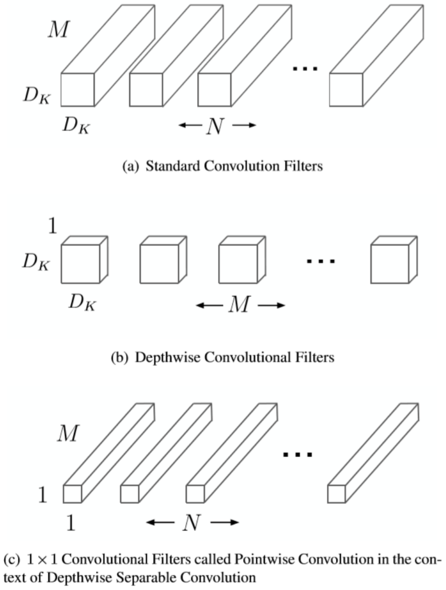
and the 2D Gaussian blur operator is given by

$$G_(m,n) = A \cdot exp(-\frac{(m-\mu_x)^2}{2\sigma_x^2} - \frac{(n-\mu_y)^2}{2\sigma_y^2})$$

As part of the loss function, Color loss between ground truth and generated image can be represented as:

$$L_{color}(X,Y) = \|X_b - Y_b\|_2^2$$

The idea behind this loss is to evaluate the difference in brightness and contrast between the images instead of considering the detailed resolution improvement. One advantage of this method is that color loss is more insensitive to small shifts between pixels as demonstrated in the figure 9. In this way, we are able to compute the loss function even without perfect alignment between patches.



(a) Standard Convolution Filters

(b) Depthwise Convolutional Filters

(c) 1 × 1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

Figure 6. Separation of convolutional layer into depthwise and pointwise layer. Depthwise layer fulfills the task of feature extraction, and pointwise layer combines the essential features. The distribution of jobs improves the training efficiency of neural networks without much loss of accuracy

DPED+ The overall architecture applies the conventional residual block structure: starting from 9×9 convolutional layer, we implement four residual blocks, each consists of two kernel size 3×3 layers alternated with one Batch-Normalization Layer. After the residual blocks, we concatenate kernel size 3×3 layers and one last kernel size 9×9 layer. All layers are followed by LeakyReLU layer with $a = 0.005$. The color loss and total variation loss are calculated from the 9×9 final layer. We also implement discriminator network to calculate the texture loss between and VGG-19 to calculate the content loss between the ground truth and generative images.

In order to accelerate our training progress since it is only the first procedure, we replaced the traditional convolutional layer with depthwise separable layer and pointwise layer introduced in MobileNet [19]. For MobileNet [19], the depthwise convolution applies a single filter to each input channel, while pointwise convolution applies a 1×1 convolution to combine the previous outputs. Essentially, a traditional convolutional layer overloads its functionality by filtering and combining the filtered outputs in a single step. However, the separated depthwise and pointwise convolution have a more precise distribution of work and outperform in the sense of efficiency. We experimentally show that this simple separation will improve the training effi-

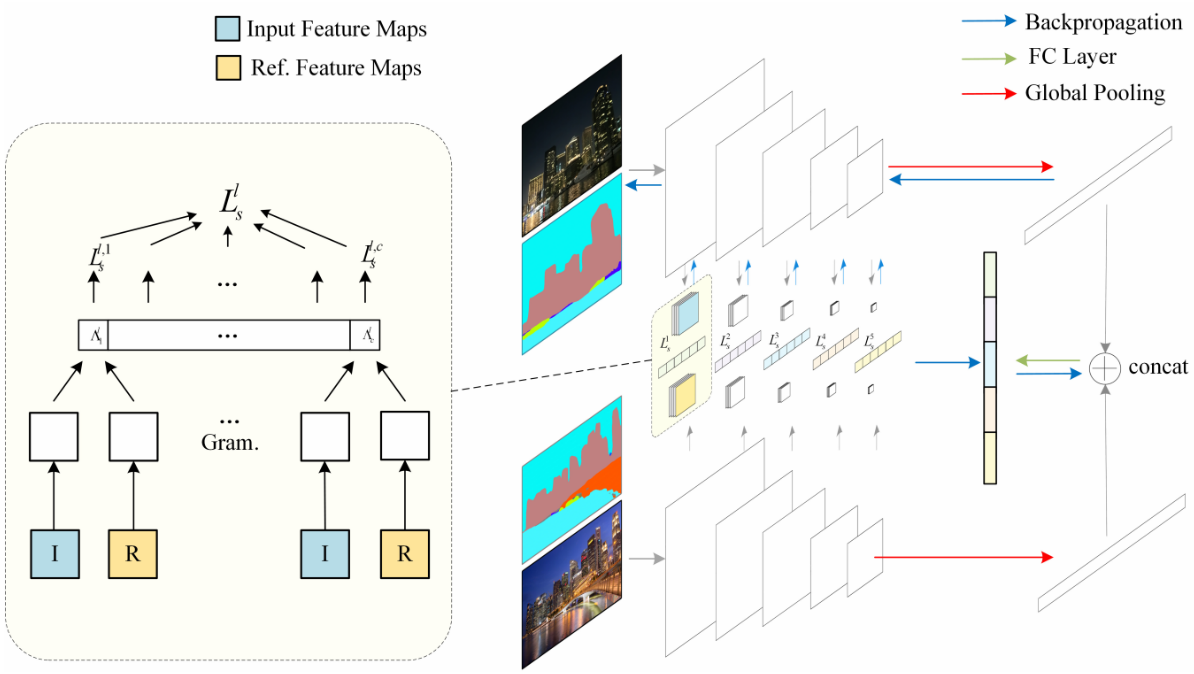


Figure 7. The network architecture of jointly enhancement factor Λ and the output image. The right side illustrates the network structure for style loss computing. We adaptively adjust our enhancement factor in regional basis and take into account the proportional information iteratively. The left side is the Gram matrix distance calculation from each region and each layer.

ciency by 40% while only suffering from PSNR degradation around 0.6 dB.

3.1.2 Semantic Style Enhancement

The second-phase enhancement takes the generated image $\mathcal{I}$ from DPED+ model as input and reference $\mathcal{R}$ as guidance to generate enhanced image $\mathcal{O}$, which is similar to the guidance image. Our framework in this part mainly consists of a night scene weakly-supervised enhancement network and a reference recommendation network.

Semantically Adaptive Enhancement The main objective of taking the style reference $\mathcal{R}$ into account is to take advantage of the photorealistics aesthetic style and perform the aesthetic adjustment automatically. Motivated by the methods of [14], we basically implement the Gram matrix to calculate the style loss in each VGG-19 layer of the image feature map. To differentiate the enhancement degrees in different regions within a single image, we adopt the semantically adaptive factor Λ and obtain the style loss:

$$\mathcal{L}_s(O,\Lambda) = \sum_{r,l} \frac{\Lambda_r^l}{2N^{l^2}} \|G_r^l[\mathcal{O}] - G_r^l[\mathcal{R}]\|_2^2$$

where r is the sematic content label for different semantic regions (i.e river, sky, building, etc), Λ_r^l represents the enhancement factor in VGG layer l and semantic region r, and N_l is the number of feature maps in layer l. $G_r^l[\mathcal{O}]$ is the Gram matrix corresponding to the feature map $F_r^l[\mathcal{O}]$ and the semantic mask $M_r^l[\mathcal{O}]$. For each segmented region of at layer l, we calculate the Euclidean distance between the generated output's and reference's Gram matrices as the regional style loss. Our total style loss of layer l is calculated as a weighted combination of all the regional style loss, which are amplified by the individual enhancement factor Λ_r^l.

The aim of our carefully designed semantic enhancement factor is to intensify the difference between the training of high-frequency details and low-frequency structures. According to our observations, different regions within the same photo always need different treatments, for instance, sky regions often lack structural details, and they normally require less attention than the texture-fruitful building regions. While most current CNN approaches ignore the regional difference and perform the image patch training on an equal basis, our method emphasizes on the discrepancies and tries to avoid the inappropriate balanced stylization.

Self-learning Enhancement Factor Λ Manual setting or hand tuning the enhancement factor is inappropriate and time-wasting. Consequently, we propose a self-learning strategy to tune the enhancement factor along with the transferring the output image to the reference style. First of all, we utilize the global pooling layer after conv5_3 to extract

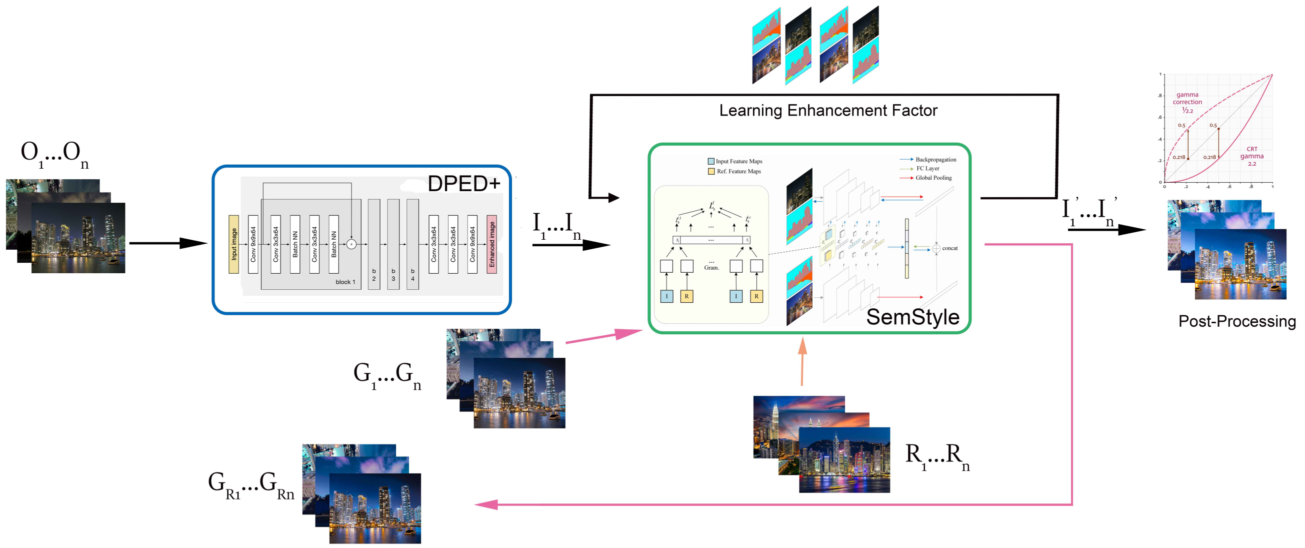


Figure 8. Pipeline of Multi-Phase Semantic Enhancement (MPSE): 1st Phase: pairwise supervised training from image $\mathcal{O}$ to the enhanced groundtruth image $\mathcal{G}_R$ using residual blocks and Generative Adversial Networks (GAN). 2nd Phase: unpaired semantic style refinement and global enhancement: Using photorealistic style transfer techniques to convert the style of image $\mathcal{I}$ to the style of reference image $\mathcal{R}$ 3rd Phase: Post-processing using Histogram Equalization and Gamma Correction to fulfill human appreciation

fixed size feature information from the generated images and reference images, which are concatenated as the input of a fully-connected layer. This fully-connected layer aims at learning the enhancement factors of different classes and different layers. Thus the output number of this FC layer would be $L \times M$, where L is the number of VGG layers and M is the number of classes in semantic segmentation. By regional heuristics we set the influential factor of the different region is proportional the area of semantic regions α_r^l. To normalize the local enhancing effect, we implement the softmax function as:

$$\Lambda_r^l = \frac{e^{\alpha_r^l H_r^l[\mathcal{O},\mathcal{R}]}}{\sum_r e^{\alpha_r^l H_r^l[\mathcal{O},\mathcal{R}]}}$$

where Λ_r^l is the semantic enhancement factor of layer l in region r, α_r^l is the regional area indicator to represent the influence of semantic region in the training process, which are calculated from the segmented content image $Seg(\mathcal{O})$ and reference image $Seg(\mathcal{R})$. H_r^l is the individual fully-connected layer output of the concatenated feature map at conv5_3 in VGG19. The whole ratio indicates the softmax function of the semantic enhancement vector.

Photorealistics Transfer Besides considering the effect of style loss, to achieve the photorealistic transfer, the texture structure and details of the original image $\mathcal{I}$ should be preserved. Consequently, we utilize a weighted combination of content loss $\mathcal{L}_c$ and photorealistic regularization loss $\mathcal{L}_m$ to maintain the local fidelity of structural information for the enhanced image $\mathcal{O}$.

$$\mathcal{L}_c(O) = \sum_l \alpha^l \frac{1}{2N^l D^l} \|F^l[\mathcal{O}] - F^l[\mathcal{I}]\|_2^2$$

$$\mathcal{L}_m(O) = \sum_{c \in r,g,b} V_c[\mathcal{O}]^T \mathcal{M}_\mathcal{I} V_c[\mathcal{O}]$$

where α^l serves as the content weight in each layer. D^l and N^l are the dimension and number of the vectorized feature map respectively. $V_c[.]$ refers to the vectorized c channel of the image, and $\mathcal{M}_\mathcal{I}$ is the linear system defined in the Matting Laplcian Matrix.

Final Objective Function and Optimization Our final objective function is a weighted combination of semantic adaptive style loss $\mathcal{L}_s$, content loss $\mathcal{L}_c$ and photorealism regularization loss $\mathcal{L}_m$:

$$\mathcal{L}(\mathcal{O}, \Lambda) = \mathcal{L}_c(\mathcal{O}) + \beta \mathcal{L}_s(\mathcal{O}, \Lambda) + \gamma \mathcal{L}_m(\mathcal{O})$$

where β is the weight of the total style loss, γ is the weight to balance the photorealism regularization loss L_m.

We apply the alternate updating scheme t optimize our final objective function. In each step, we fix one value and employ the stochastic gradient descent to update another. We optimize the output image by computing $\frac{\partial \mathcal{L}}{\partial \mathcal{O}}$ in the backpropagation procedure in the total loss. As for the enhancement factors, we update the gradients of fully connected layer as follows:

$$\frac{\partial \mathcal{L}}{\partial W_{ci+j,k}} = \sum_{r,l} \frac{\partial \mathcal{L}}{\partial \Lambda_{r]^l}} \frac{\Lambda_r^l}{\partial W \alpha_r^l H_r^l[\mathcal{O},\mathcal{R}]_{i,j}} \frac{\partial W \alpha_r^l H_r^l[\mathcal{O},\mathcal{R}]_{i,j}}{W_{ci+j,k}}$$

$$= \sum_{r,l} \mathcal{L}_s^{r,l} \alpha r^l (\delta_{i,l}\delta_{j,r}\Lambda_r^l - \Lambda_r^l \Lambda_j^i) H_r^l[\mathcal{O},\mathcal{R}]_k$$

where L_s is total style loss, c is the number of semantic classes, (i, j) states it is the i-th layer and j-th class, k

shows that it is the k-th element of the global pooling feature vector. $ci + j$ is the row, and k is the column of fully connected layer's parameter matrix.

Reference Recommendation Network The selection of appropriate reference images is crucial in night-view image enhancement. We selected up to 4000 professional night-view night photographs from image sharing websites, such as Flicker. We basically select 5 axiomatic semantics from the websites such as sky, building, river, road, and vehicles. A variety of shooting positions and styles (sky view, panoramic,etc.) are considered to diversify our proposed dataset. In order to retrieve the appropriate reference style image from the large dataset, we first define the semantic metric D_{sem} as the Euclidean distance of features in FC-8 layer of pre-trained VGG-16 classification network, the semantic similarity is defined as the following:

$$D_{sem} = \|f_{fc8}(I_1 - f_{fc8}(I_1)\|_2$$

where f_{fc8} defines the FC-8 features of an image. For an original image x_i, we find best 60 semantically similar images in the dataset for the further selection. We not only requires the close relationship between semantic contexts of two images, but also an appropriate artistic style. By selecting 10 of the 60 candidate samples, we label them as positive references $x^{p_1}, x^{p_2}, ..., x^{p_{10}}$ and the rest as negative references $x^{n_1}, x^{n_2}, ..., x^{n_{50}}$. Our goal is to minimize the distances between the anchor x and positive set and maximize the distances between x and negative set:

$$\mathcal{L}_{ref} = \arg\min_{\mathcal{L}}[\sum_{j=1}^{10} \|f_(x_i) - f_(x_i^{p_j})\|_2 - \sum_{k=1}^{50} \|f_(x_i) - f_(x_i^{n_k})\|_2 + \alpha]_\dagger$$

where α is a margin to enforce a minimum distance between positive and negative samples, f is the output feature from our recommendation network, and $[.]_\dagger$ denotes $ReLU$ function.

To build a simple structural recommendation system, we additionally include a fully-connected layer after FC-8 layer of the pre-trained VGG-16 network. Only the parameters of the extra fully-connected layers are tuned, and previous parameters are all fixed during the training stage. After training our images upon the artistic night-view dataset, our reference recommendation network will automatically provide the corresponding appropriate reference image for an input image unseen before.

3.1.3 Post-processing Optimization

After the deep neural network training of the original images, we further investigated two traditional image processing methods(Histogram Equalization, Gamma Correction) to improve our image qualities further.

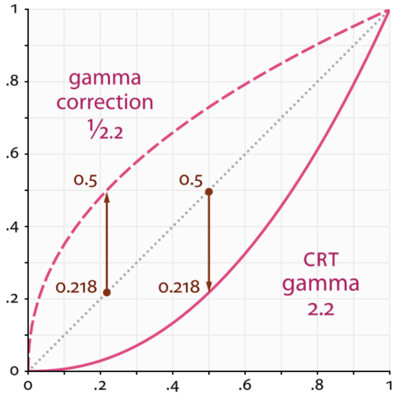


Figure 9. Comparison between responses to different light intensity of human eye (Gamma 1/2.2) and commonly-used cameras (Linear).

Histogram Equalization Histogram equalization is one of the efficient image enhancement method to adjust the contrast. Based on our previously computed regional influential factor α_r^l and enhancement factor Λ_r^l, we are able to perform the image decomposition recursively and generate sub-images of different semantic regions. By adapting the cdf (cumulative density function) to the original histogram in each sub-image, we can convert the image histogram into a uniform distribution and maximize the information entropy in each semantic area:

$$T(k) = floor(L-1)\sum_{n=0}^{k} p_n$$

$$\frac{d}{dy}(\int_0^y p_Y(z)dz) = p_x(T^{-1}(y))\frac{d}{dy}(T^{-1}(y))$$

$$\frac{d}{dx}\Big|_{x=T^{-1}(y)}\frac{d}{dy}(T^{-1}(y)) = 1$$

which means $p_Y(y) = \frac{1}{L-1}$, HE flattens the histogram in individual semantic region and improve the contrast.

Gamma Correction Based on the Weber-Fechner's Law, we discovered that the response of human eyes to different illuminance conditions follow a logarithm curve and react stronger to low-light condition than the strong-light condition. Consequently, when we take the physical reaction of human being into account, we proposed to follow the logarithm curve and adjust it using gamma correction $1/2.2$ rather than linear, which serves as the last procedure of our first algorithm.

3.2. Multi-Domain Intrinsic Enhancement

We describe the second model for image enhancement as a 'Decompose-Enhancement-Reconstruct' pipeline. In this part, we first discuss a self-supervised framework to obtain the intrinsic components of real-life images, then we

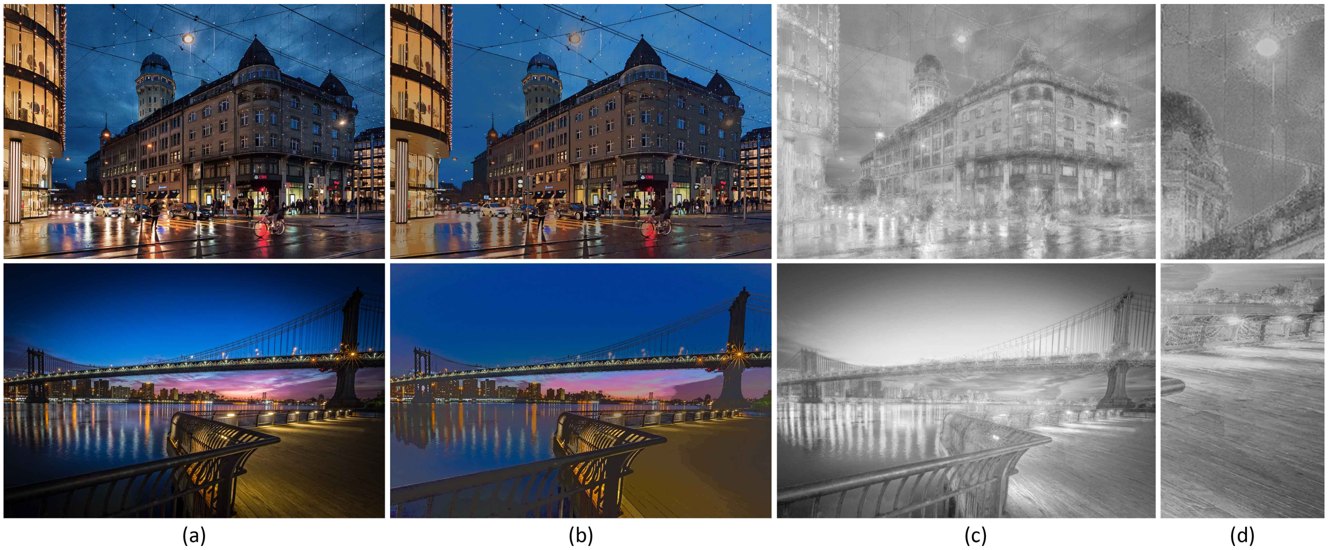


Figure 10. Intrinsic decomposition of the mobile and DSLR image: (a) is the RGB image, (b) is the reflectance map, (c) is the shading, (d) is a zoom-in observation of the shading. The first row is the mobile image; the second raw is the image taken from DSLR. We can observe that the reflectance map largely affects the color and contrast of the image whereas the shading affects the texture and noise aspects of the image. A zoom-in sample (d) demonstrates the concentration of noise in the shading.

discuss a specialized GAN [10] architecture which handles multi-domain image translation for overall image enhancement. Later, the final enhanced image is constructed by an intrinsic composition framework, which utilizes the output of GAN introduced. In the last, we show that all components can be trained and fine-tuned together as an automatic three-step framework.

3.2.1 Self-Supervised Intrinsic Decomposition

In intrinsic decomposition, image $\mathcal{I}$ is described as the point-wise multiplication of the reflectance map $\mathcal{A}$ and shading $\mathcal{S}$.

$$\mathcal{I} = \mathcal{A} \cdot \mathcal{S}$$

In some paper, the reflectance map $\mathcal{A}$ is also called Albeto. Until now, there exists merely work attempting to improve the image quality by manipulating its intrinsic decomposition. In our experiment of splitting mobile and DSLR image into Albeto and shading, we find that the effect of color and contrast is centralized in Albeto, whereas the impact of noise and texture is centralized in the shading. The experiment result is shown in figure 10.

We have the following observation: In daylight condition, the ambient light is relatively abundant, and the dominating difference of mobile phone and DSLR is mainly revealed in Albeto. In the night-scene situation, with poor lighting condition, the dominating difference of mobile phone and DSLR is revealed in shading. This assumption allows us to explore image enhancement from different aspects by finding its intrinsic components.

Reconstruction Loss Before our attempt, finding the possible intrinsic components for images has been extensively studied. Most data-driven methods rely solely on ground truth labelling, and because hand-labeling Albeto and shading are nearly impossible to achieve, most methods still rely on computer-generated virtual data like MPI Sintel Dataset [5]. As a result, those methods perform not well in real-life photos. The previous model assumes access to ground truth labels for all inputs and does not explicitly model the reconstruction of the input image based on intrinsic image predictions.

To alleviate that problem, the intrinsic model should be trained with real data, and we introduce a network with self-supervision proposed by [23] to find real-life intrinsics. The basic idea is to reconstruct the intrinsic predictions back to an RGB image, and the image should look like the input RGB image as close as possible. If $\mathcal{F}$ is the algorithm finding the intrinsic with

$$\mathcal{F}(\mathcal{I}) = (\mathcal{A}, \mathcal{S})$$

Then we aim to minimize:

$$\|\mathcal{I} - \mathcal{A} \cdot \mathcal{S}\|_2^2$$

Despite the intention is quite reasonable, directly minimizing the reconstruction loss could be erroneous. A simpler meaningless solution that yields zero reconstruction is:

$$\mathcal{A} = \mathcal{I}, \mathcal{S} = 11^T$$

where 1 is the matrix with all elements to be 1. This indicates the necessity of substituting the reconstruction dot

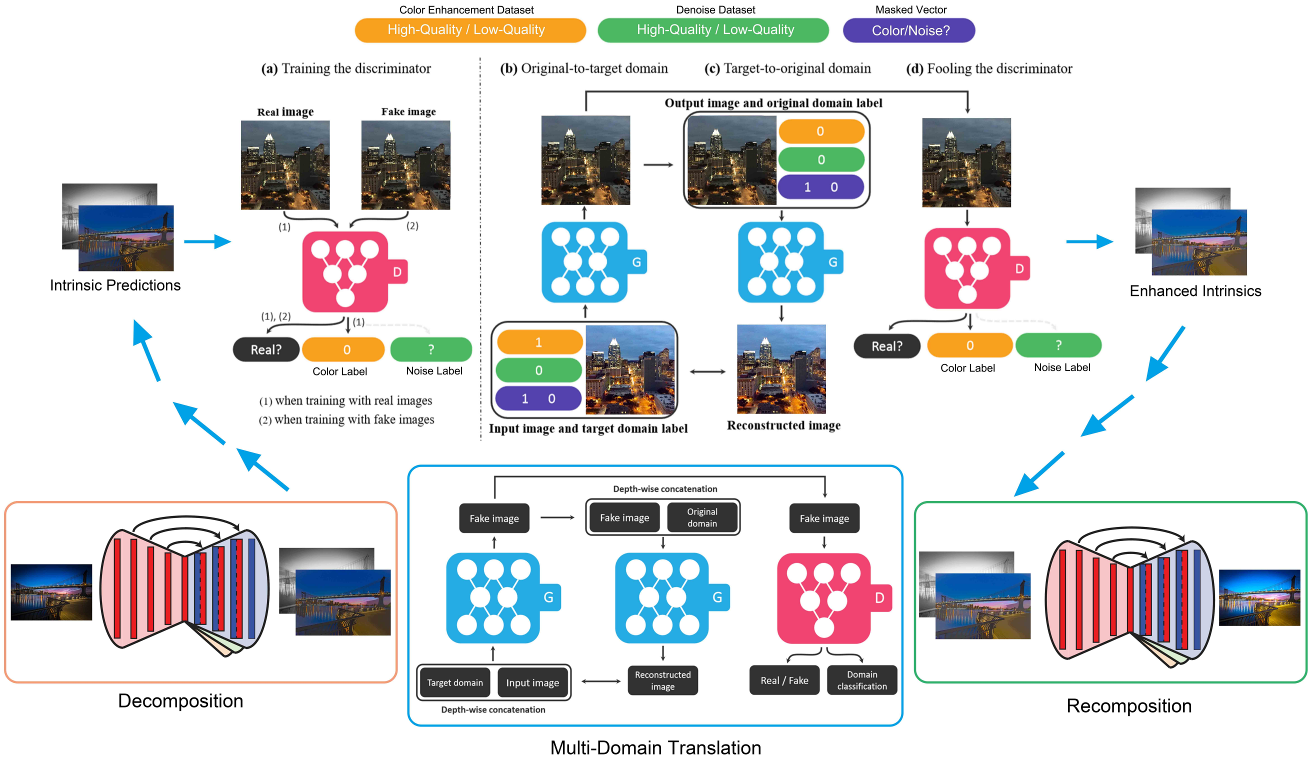


Figure 11. Network Architecture of Multi-Domain Intrinsic Decomposition (MDIE): The overall network finds the input image's intrinsic decomposition, then concat the albeto and shading into a six-channel input and pass it the multi-domain GAN in the second. The GAN then utilize several datasets specified for partial enhancement such as color, noise, and texture, by the help of a masked vector, we expect the final generator can output a comprehensively enhanced image when we name all label to 'high-quality'. After the multi-domain GAN, we use another intrinsic network to reconstruct the intrinsics back to an enhancement image. The reconstruction loss is also calculated in the last step.

product by another composition algorithm $\mathcal{F}'$ so that the reconstruction loss is:

$$\mathcal{L}_{rec} = \|\mathcal{I} - \mathcal{F}'(\mathcal{A}, \mathcal{S})\|^2$$

This provides our intrinsic prediction framework a possible source to learn an approximation of the exact decomposition from unlabelled data.

Symmetric Intrinsic Network In our design, we set both the decomposition framework $\mathcal{F}$ and composition framework $\mathcal{F}'$ to be an autoencoder-autodecoder architecture. The encoder has 5 convolutional layer with $\{16,32,64,128,256\}$ filters of size 3×3 and stride 2. Batch normalization and ReLU is used in every layer. The decoder has the same architecture with the reversed order, with finally an output layer with six output channels dedicated to Albeto and ground truth. Sub-Pixel upsampling [36] is used after the convolutional layer to enlarge the feature maps and speed up the upsample operation, which is proved to be much faster than the deconvolutional layer [32].

3.2.2 Multi-Domain Image-to-Image Translation

The task of image enhancement can be thought as a process of image-to-image translation, where we expect the translation is from a 'mobile-quality' image to a 'DSLR-quality' image. Intuitively, we can define the former set as 'low-quality' domain and 'high-quality' domain and find a translation between the two domain. However, this attempt suffers from two potential obstacles:

- First, training a dual-side image-to-image translation network like [2, 21] requires pair-wise data, where the image in 'mobile-quality' and 'DSLR-quality' should be precisely aligned. In reality, even if the two images are taken in the same scene and with the same setting, image registration is hard to achieve due to different optics and distortions.
- Second, training directions is unclear with hidden contributing factors behind 'high-quality'. Usually, those factors are mixed, including high dynamic range, vibrant colors, low noise level, etc. Simply training the network in end-to-end fashion will inevitably cause ambiguities in enhance directions. For instance, the noise pattern might be indistinguishable with some structural information, which leads to the appearance of noise and loss of some structural information in the final results.

To avoid these two obstacles, we utilize datasets dedicated to the individual factor of image enhancement and propose a vector-mask GAN architecture to achieve image-to-image translation in multiple domains. By this way, we can handle the problem absence of one overall enhancement dataset with perfect alignment. Because for one time we only train the network with a specific enhancement factor with a mask vector, the problem of ambiguity in training direction can be alleviated.

We denote the term *attribute* as a meaningful enhancement factor inherent in an image such as color, texture, or noise level and *attribute value* as a particular level of the attribute, such as low-quality or high-quality. We further denote *domain* as a set of images sharing the same attribute value. For example, images of low-quality in color perspective can be one domain whereas images of high-quality in color perspective can be another domain.

Our goal is to train a single generator G that learns the mappings among the three domains we specified: color, texture, and noise. To achieve this, we train G to translate an input image x into an output image y conditioned on the target domain label c, $G(x, c) \to y$. We randomly generate samples from the three domains and train G from three different enhancement factors. We also introduce an auxiliary classifier [33] that allow a single discriminator to control multiple domains. In that way, our discriminator produces probability distributions over both sources and domain labels, $D : x \to \{D_{src}(x), D_{cls}(x)\}$.

Mask Vector To handle the lack of some labels when training sample in one domain, we introduce a mask vector m to let the GAN explicitly ignore unknown dataset and only concentrates on one specific dataset. We use an n-dimensional one-hot vector to denotes m with n being the number of datasets we used. In addition, we define the label c as a vector in a unified version:

$$\widetilde{c} = [c_1, ., c_i, ., c_n, m]$$

If here we input the i-th dataset, then only c_i is assigned to 1 and others is assigned to 0. In our experiments, we have three datasets dedicated to color, texture, and noise. Then the n is set to be 3. In each dataset, we only have two attribute value: low-quality and high quality.

Adversarial Loss Similar to previous GAN design, to ensure the images in one specific domain indistinguishable from real images in that domain, we define the adversarial loss as:

$$\mathcal{L}_{adv} = \mathbb{E}[logD_{src}(x)] + \mathbb{E}[log(1 - D_{src}(G(x, c)))]$$

where G generates an image $G(x, c)$ conditional on the input x and the target domain label c, whereas D is designed to distinguish between real and fake high-quality image. We further denote the term $D_{src}(x)$ as the probability distribution over sources given by D. In this setting, the generator D aims to minimize the objective, while discriminator D tries to maximize the objective.

Domain Classification Loss For a given input image x and a target domain label c, our goal is to translate x into an output image y, which is appropriately classified into the target domain c. To satisfy this requirement, an auxiliary classifier is added on top of the discriminator D and the domain classification loss is imposed when optimizing both the discriminator D and generator G. We use the domain classification loss of real images to optimize D:

$$\mathcal{L}^r_{cls} = \mathbb{E}_{x,c'}[-logD_{cls}(c'|x)]$$

where the term $D_{cls}(c'|x)$ denotes a probability distribution over domain labels calculate by D. By minimizing this objective, D learns to classify a real image x to its corresponding original domain c'. Then we use domain classification loss of fake images to optimize G:

$$\mathcal{L}^f_{cls} = \mathbb{E}_{x,c}[-logD_{cls}(c|G(x, c))]$$

where G attempts to minimizing this objective to generate images that can be classified into the target domain c.

Cycle Reconstruction Loss Consider our task handles a multi-domain problem, minimizing the loss mentioned above does not guarantee the translated image preserve the enhancement factors of its input images while only changing domain-related part of inputs. To address this potential problem, we introduce a cycle consistency loss [45] defined as:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x,c,c'}[\|x - G(G(x, c), c')\|]_1$$

where G takes the translated image $G(x, c)$ and the domain label of the original input c' as input and attempts to reconstruct the original image x.

Full Objective Optimization The full objective function to optimize generator G and discriminator D is:

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^r_{cls}$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^f_{cls} + \lambda_{cyc}\mathcal{L}_{cyc}$$

where λ_{cls} and λ_{cyc} are hyper-parameters that balance the relative weight of domain classification loss and cycle reconstruction loss. In our experiment, λ_{cls} is set to be 1 and λ_{cyc} is set to be 10.

3.2.3 MDIE Ojective Optimization

We consider the loss proposed in the intrinsic part and multi-domain translation part when we train the network altogether, α and β are also the weight factor balancing the two reconstruction loss to stabilize the convergence of learning to predict the intrinsics:

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^r_{cls}$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^f_{cls} + \lambda_{cyc}\mathcal{L}_{cyc}$$

$$\mathcal{L}_{rec} = \alpha\|\mathcal{I}_{LQ} - \mathcal{F}'(\mathcal{A}_{LQ}, \mathcal{S}_{LQ})\|^2 + \beta\|\mathcal{I}_{HQ} - \mathcal{F}'(\mathcal{A}_{HQ}, \mathcal{S}_{HQ})\|^2$$

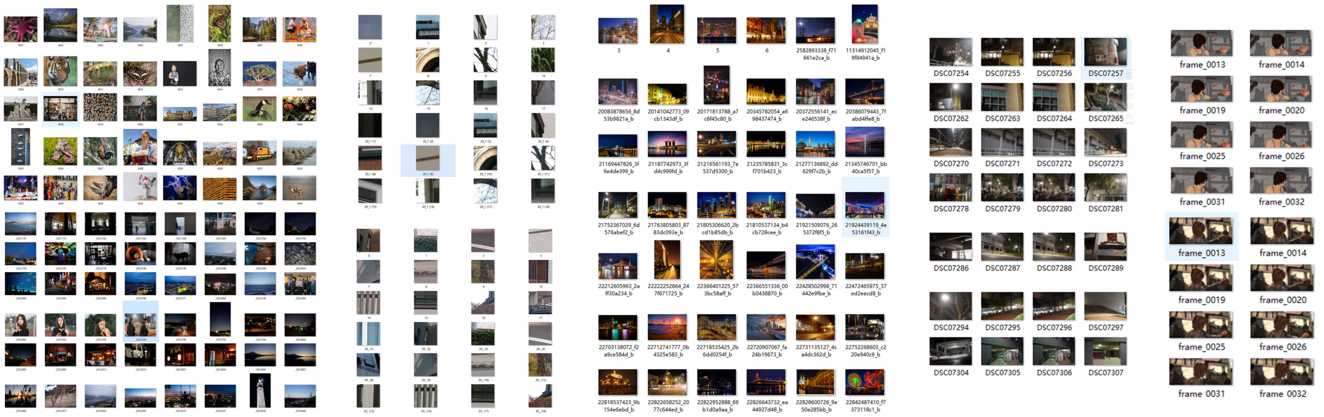


Figure 12. Overview of Datasets: From left to right is: the Enriched DIV2K Dataset, DPED PR Dataset, Mobile-DSLR UPR Dataset, Damstadt Noise Dataset, MPI Sintel Dataset.

4. Dataset

4.1. DIV2K Dataset

Inspired by the NTIRE 2017 Challenge [38] on Single Image Super-Resolution competition, we obtained the newly proposed high-quality data-set DIV2K for our tasks. Consisting of 800 training images, 100 validation images and 100 test images shot by DSLR cameras, DIV2K dataset outperformed most existing benchmark datasets and achieved excellent training results. Considering DIV2K dataset only consists of images under daylight, we especially add 500 night images of various low-light condition shot by our DSLR.

4.2. PR (Paired) Dataset

Apart from the DIV2K dataset which specified for one sub-domain of image enhancement task, we also considered the general image quality assessment for our training as mentioned earlier method, in whose case we are aiming at collecting paired image set without the restrictions of perfect alignment.Apart from the existing DSLR Photo Enhancement Dataset (DPED) at hand, we collected over 800 photos (400 pairs) for specific night-view enhancement tasks. In each pair, the high-quality image is a 24-megapixel image shut by our full-frame DSLR, equipped with six stops of exposure to achieve higher dynamic range; whereas the low-quality image is a 12-megapixel image shut by our iPhone6S, without any further adjustments. All the photography devices were mounted on the tripod and controlled by a shared shutter, which ensures paired images have been shot at the same time. During the shooting procedures, we adjusted digital parameters and device settings and tried our best to maximize the alignment ratio between paired images. Together our PR dataset consists of over 2K image pairs, with 60% being the day view pairs coming from DPED dataset, and 40% being the night view pairs coming from our contribution.

4.3. UPR (Unpaired) Dataset

One benefit of using unpaired framework for training is the profusion of vast images of different qualities on the Internet. More than reusing the previous images in an unpaired manner, we claw more high-quality images shot by professional photographers from Flickr Spotlight API. Those images are not only rich in texture and detail but also superior in the lighting condition, color balancing and structure after professional human retouch. The UPR dataset we collect contains two parts: the low-quality part contains over 500 images shot by our iPhone, and the high-quality part contains over 2000 professional Flickr images.

4.4. Damstadt Noise Dataset

The lack of the ground truth data causes the benchmark of the denoise task relying on synthesized i.i.d. Gaussian noise. This approach can be problematic, considering noise in real images far from i.i.d. Gaussian. The work [35] presents a novel denoising dataset called the Darmstadt Noise Dataset (DND). It consists of 50 pairs of real noisy images and corresponding ground truth images that were captured with consumer grade cameras of different sensor sizes. For each pair, a reference image is taken with the base ISO level while the noisy image is taken with higher ISO and appropriately adjusted exposure time. The reference image undergoes a careful post-processing entailing small camera shift adjustment, linear intensity scaling, and removal of low-frequency bias.

Additionally, considering the noise pattern is different for different sensors such as mobile phones and DSLR cameras, constructing a specialized denoising dataset is essential. Here we sample data under different ISO setting from Mobile Phone and DSLR, then use the post-processing techniques in [35] to approximate the denoised ground truth.

4.5. MPI Sintel Dataset

The MPI Sintel dataset is initially proposed in [5] for optical-flow evaluation, where it provides depth image, disparity maps, optical flow maps, stereo images, and albeto in an animated movie. We use the albeto to infer the ground truth shading, and use it to pre-train the first intrinsic decomposition VAE network.

5. Experiment

5.1. MPSE Experiment

Although we have set our objective to improving mobile phone image qualities generally by deep learning methods. we conducted our experiments especially on low-light images since under night-view situations, images shot by mobile devices are more likely to exhibit problems like lacking details, high noise, and low dynamic range. The last term we have completed most state-of-the-art Super-Resolution deep neural networks and explored their advantages and disadvantages. Consequently for this term, when employing our methods to low-light image enhancement tasks, we can still make use of the benchmarks and training strategies to maximize our performances.

Convolutional Neural Network Models				
Iterations	2×10^4	2×10^4	7×10^4	7×10^4
Revised Models	PSNR	SSIM	PSNR	SSIM
Baseline	26.3455	0.9112	27.8025	0.9338
VGG Loss(VL)	27.4288	0.9493	28.8042	0.9703
$l1 + l2$	26.2453	0.8954	26.8251	0.9122
MobileNet	25.2410	0.9321	27.4123	0.9416
LReLU(LR)	26.5448	0.9107	27.7239	0.9293
Contrast(HE)	27.0145	0.9220	28.0112	0.9445
VL+LR+HE	**27.4338**	**0.9420**	**29.2002**	**0.9754**

DPED+ Phase According to our observation in the benchmark experiment above, we notice that leaky ReLU, MobileNet and VGG loss all help boost the network performance, so we implement our pair-wise training network using the similar techniques. In the actual experiments, we find that Leaky ReLU helps the network converge faster so that there is no need for extra iterative training, MobileNet helps divide the feature extraction and combination tasks and improve the training accuracy, VGG content loss augment extra regularizations and keep the fidelity of the original image. For the consideration of overall efficiency, we choose the Nvidia Titan X GPU for 30K iterations, setting the batch size 50. The parameters of the neural network were optimized using Adam optimizer for gradient descent with learning rate 5e-4. In every 10K iterations, we cut down the learning rate to the half of the original one. All the training curves are recorded in the Tensorboard, and we can see a promising trend of convergence.

Semantic Style Enhancement In order to retrieve the style image from the original one, we incorporate with PSPNet with *ADK20* dataset to perform the semantic segmentation and merge the similar contents into the same semantic labels. We select the balancing weights as $\{\beta, \gamma\} = \{100, 11\}$ as heuristics. The initialization of parameters is aiming at balancing the scale of different losses to make them effective. In the actual experimental training, we augment the original loss function with total variation loss $\mathcal{L}_{tv}$ to make the convergence direction towards sharp edges and sparse representations.

Post-Processing Optimization In order to optimize the effect of image enhancement and make it suitable for artistic appreciation, we propose the Recursively Histogram Equalization (RHE) method according to the enhancement factor Λ in conv5_3 and gamma correction at ratio $1/2.2$. Comparing to the CycleGAN, DeepTransfer and Deep Analogy method, our method shows not only significant improvement in structural details and overall contrast but also exhibits promising result in the perceptual user study.

Perceptual User Survey Based on the subjective nature of image enhancement tasks, we additionally conducted a perceptual user survey to evaluate our performances of pipelines. By recruiting 40 professional photographers and artists, we provided them with 15 sets of images; each contains seven night-view photos including original image, Histogram Equalization, Cycle-GAN, WESPE DeepTransfer, Deep Analogy and MPSE (our approach). The users are asked to rank them according to personal preference and artistic taste. The detailed survey results are listed below.

Night-View Image Enhancer Models				
Night-view Models	PSNR	SSIM	1st Choice	2nd Choice
Input	26.3455	0.8112	0.0	0.6
HE(Contrast)	26.3421	0.8453	5.4	4.0
Cycle-GAN	27.4390	0.9239	18.2	16.8
WESPE	26.8442	0.8902	12.0	10.4
Deep-Transfer	27.5104	0.9177	25.0	21.2
Deep Analogy	28.0130	0.9212	19.0	20.2
MPSE(Ours)	**28.4128**	**0.9544**	**30.4**	**26.8**

It suffices to show that our performance metrics exhibit great improvements regarding PSNR and SSIM. Although the user perceptual evaluation is subjective to individual perspectives, we still inspect considerable improvements over the current state-of-the-art Deep Photorealistic Transfer techniques by roughly 5%.

5.2. MDIE Experiment

Pre-train Intrinsic Network Although we design our intrinsic prediction network in a self-supervised fashion, the actual training will still need a pre-train phase with albeto and shading as the ground truth. Here we first pre-train the two VAEs in our intrinsic network with MPI Sintel Dataset. The loss in albeto, shading and reconstruction will be rebalanced in the pre-trained phase.

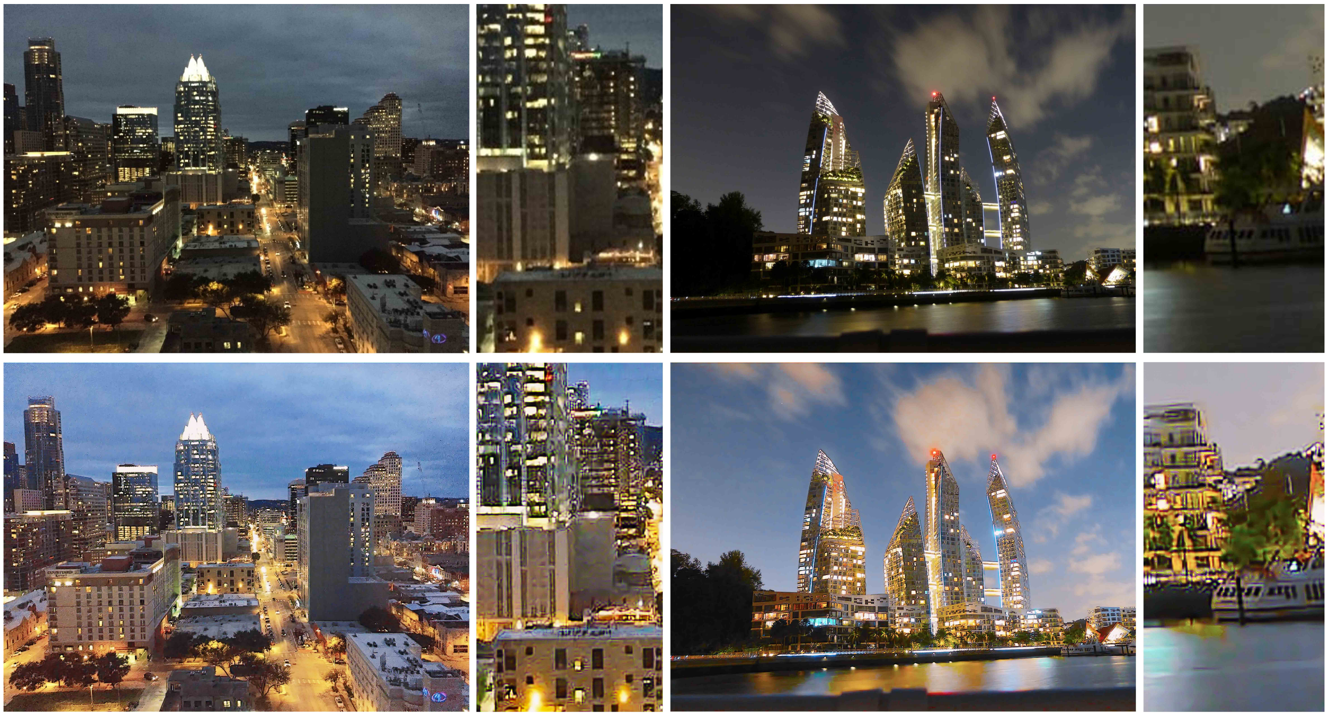


Figure 13. Effect of the whole MPSE pipeline: In combination with DPED+ and post-processing techniques, it can be observed that the illuminance condition has been improved and our pipeline has certain capabilities to improve the structure information like edges.

Training MDIE In this step, we use the weight of intrinsic network in the pre-train step, and this time we train the whole network in a combined fashion. For the multi-domain translation GAN, we have three datasets concerning color, noise and texture enhancement respectively:

- Color Enhancement Dataset: We collect result created by professional photographers' human-retouching techniques. This dataset contains the original RAW file and the retouched image outputted as JPEG. The color quality and contrast condition of the retouched image is significantly improved.
- Denoise Dataset: We collect noise image of the same scene from Mobile-Phones, with different ISO settings. Then we use the algorithm introduced in [35] to approximate the ground truth image with zero noise.
- Texture Enhancement Dataset: We collect high-quality night scene images from DSLR camera, and then we downgraded these images and created their 'low-quality' counterparts.This dataset is used to enhance the image from texture and structural perspective partially.

6. Conclusion

Aiming at improving the low-light mobile image qualities by deep neural networks, our thesis proposed two systematic pipelines: Our first enhancement algorithm **MPSE** (Multi-Phase Semantic Enhancement) initially trains an end-to-end paired residual model to augment the structural details and texture information of images, then performs semantic style refinement while keeping the fidelity of the generated image, finally utilizes post-processing techniques like Histogram Equalization and Gamma Correction to fulfill human artistic appreciation. The second method **MDIE** (Multi-Domain Intrinsic Enhancement) introduces the usage of intrinsic decomposition in the task of image enhancement. To facilitate the difficulty of lacking precisely-aligned Mobile-DSLR dataset, the second method designs a multi-domain image-to-image translation network to learn a comprehensive enhancement transformation from partially enhancement dataset. Both of our methods attempt to adopt mid-level vision ideas to low-level vision and provides better supervision in semantic or intrinsic perspective.

References

[1] M. E. A. Selim and L. Doyle. Painting style transfer for head portraits using convolutional neural networks. *TOG*, 2016.

[2] R. T. K. V. Andrey Ignatov, Nikolay Kobyshev and L. V. Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *IEEE International Conference on Computer Vision (ICCV), 2017*, 2017.

[3] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Trans. Graph.*, 2014.

[4] H. C. Burger, C. J. Schuler, and S. Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012.

[6] H. Chang, D.-Y. Yeung, and Y. Xiong. Super-resolution through neighbor embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[7] K. H. Chao Dong, Chen Change Loy and X. Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[8] X. T. Chao Dong, Chen Change Loy. Accelerating the super-resolution convolutional neural network. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[9] P. Chatterjee and P. Milanfar. Clustering-based denoising with locally learned dictionaries. *IEEE Transactions on Image Processing*, 2009.

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017.

[11] K. Dabov, A. Foi, and K. Egiazarian. Video denoising by sparse 3d transform-domain collaborative filtering. In *European Signal Processing Conference*, 2007.

[12] B. G. E. Reinhard, M. Adhikhmin and P. Shirley. Color transfer between images. *IEEE*, 2001.

[13] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 1999.

[14] E. S. K. B. Fujun Luan, Sylvain Paris. Deep photo style transfer. 2017.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*. 2014.

[16] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision*, 2009.

[17] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang. Convolutional sparse coding for image super-resolution. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[18] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, 2001.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[20] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. V. Gool. WESPE: weakly supervised photo enhancer for digital cameras. *CoRR*, abs/1709.01118, 2017.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Y. L. P. U. K. Q. W. K. B. J. R. Gardner, M. J. Kusner and J. E. Hopcrof. Deep manifold traversal: Changing labels with convolutional features. *CoRR*, 2015.

[23] M. Janner, J. Wu, T. Kulkarni, I. Yildirim, and J. B. Tenenbaum. Self-Supervised Intrinsic Image Decomposition. In *Advances In Neural Information Processing Systems*, 2017.

[24] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] T. Khler, X. Huang, F. Schebesch, A. Aichert, A. Maier, and J. Hornegger. Robust multiframe super-resolution employing iteratively re-weighted minimization. *IEEE Transactions on Computational Imaging*, 2016.

[26] K. Li, Y. Zhu, J. Yang, and J. Jiang. Video super-resolution using an adaptive superpixel-guided auto-regressive model. *Pattern Recognition*, 2016.

[27] X. Li and M. T. Orchard. New edge-directed interpolation. *IEEE Transactions on Image Processing*, 2001.

[28] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[29] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[30] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] X. Mao, C. Shen, and Y. Yang. Image restoration using convolutional auto-encoders with symmetric skip connections. *CoRR*, abs/1606.08921, 2016.

[32] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

[33] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017.

[34] S. S. Omprakash Patel, Yogendra P. S. Maravi. A comparative study of histogram equalization based image enhancement techniques for brightness preservation and contrast enhancement. 2013.

[35] T. Plotz and S. Roth. Benchmarking denoising algorithms with real photographs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[36] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[37] J. Sun, Z. Xu, and H.-Y. Shum. Image super-resolution using gradient profile prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[38] R. Timofte, E. Agustsson, L. V. Gool, and Others. Ntire 2017 challenge on single image super-resolution: Methods and results. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[39] P. Vincent, H. Larochelle, Y. Bengio, and P. antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[40] Z. Wang, Y. Yang, Z. Wang, S. Chang, J. Yang, and T. S. Huang. Learning super-resolution jointly from external and internal examples. *IEEE Transactions on Image Processing*, 2015.

[41] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, 2012.

[42] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 2010.

[43] L. Zhang and X. Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Transactions on Image Processing*, 2006.

[44] Q. Zhou, S. Chen, J. Liu, and X. Tang. Edge-preserving single image super-resolution. In *Proceedings of the 19th ACM International Conference on Multimedia (MM)*, 2011.

[45] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.