WESPE: Weakly Supervised Photo Enhancer for Digital Cameras

Felix Singerman - 7970742, Rishabh Kukreja - 300086824

Andrey Ignatov, Nikolay Kobyshev, Kenneth Vanhoey, Radu Timofte, Luc Van Gool

Abstract

Low end devices lack the capabilities to produce a high-quality image due to hardware and software constraints. In this paper, the aim was to improve the quality of images taken by low end mobile and convert it into a high-quality image. The approach is based on weak supervised learning i.e. without the to enhance the image quality without the requirement of the ground truth. The paper presents a novel image-to-image Generative Adversarial Network (GAN) based architecture trained under weak supervision in order to mitigate the limitations of poor photo quality from mobile cameras and old photos.

1. Introduction

The boom in the technology has led to the development of smartphones that are able to capture the unprecedented detail and color. The past of these smartphones and cameras, they lack the advancements and produced poor quality images which look dull and unimaginative. Image enhancement can be performed by graphical artists or by using specialized software that enhances the photo's sharpness, contrast adjustments, and much more. However, this requires a set of skills that the day-to-day user might not have. Additionally, the process can be lengthy, and it is not possible to compute in a reasonable amount of time in the case of large-scale data processing. Past approaches to this problem have been through a supervised learning method which requires matched before/after learning pairs.

WESPE[1] was proposed at the Conference of Computer Vision and Pattern Recognition (CVPR) in 2018, by Andrey Ignatov, Nikolay Kobyshev, Kenneth Vanhoey, Radu Timofte, and Luc Van Gool from ETH University. WESPE presents a novel image-to-image Generative Adversarial Network (GAN) based architecture trained under weak supervision in order to mitigate the limitations of poor photo quality from mobile cameras and old photos.

Many of the previous methods were fully-supervised putting the requirement of having a matched pair of before and after images, thus requiring a large amount of work in order to get the matched pair and causing a deficiency in the color and texture transfer for the photo enhancement. WESPE eliminates this requirement and deficiency by being trained with weak supervision thus eliminating the need to pixel-aligned image pairs. This allows WESPE to be repeatable on virtually any camera while achieving comparable or superior results as past methods.

Figure 1: Cityscapes image enhanced

2. Dataset

WESPE only requires two distinct datasets, one from the origin camera that we would like to improve and one composed of high-quality images. The images in these datasets are all of different sizes but all of these

contains 3 channels. These 3 channels are RGB channels that correspond to colored images.

For our model, we'll be using the publicly available dataset DPED [2], Cityscapes [3], and Kitti [4]. The DPED dataset contains images from 3 smartphones from low to middle-end cameras (iPhone 3GS, Blackberry Passport, Sony Xperia Z) paired with images of the same scene taken by a high-end camera.

Cityscapes and Kitti contains a large number of urban-images of low quality, which are good candidates for automated photo enhancement. Cityscapes contain images taken by a dash-camera, so the images lack details, brightness, and resolution. While the images in the Kitti dataset are brighter but are of half the resolution as a result images lack the sharp details. For high-quality images and diverse contents, we are using the DIV2K dataset [5].

Camera source	Sensor	Image size	Photo quality	train images
iPhone 3GS	3MP	2048 × 1536	Poor	5614
BlackBerry Passport	13MP	4160 × 3120	Mediocre	5902
Sony Xperia Z	13MP	2592 × 1944	Good	4427
Canon 70D DSLR	20MP	3648 × 2432	Excellent	5902

Table 1: DPED Dataset

Camera Source	Sensor	Image size	Photo quality
KITTI	N/A	1392× 512	Poor
Cityscapes	N/A	2048 × 1024	Poor
HTC One M9	20MP	5376 × 3752	Good
Huawei P9	12MP	3968 × 2976	Good
iPhone 6	8MP	3264 × 2448	Good
FSS	N/A	>1600 × 1200	Poor-Excellent
DIV2K	N/A	~2040 × 1500	Excellent

Table 2: Additional Dataset for WESPE

3. Algorithm and Implementation

For our implementation, we have decided to code the GAN of the network. We use the GAN that is proposed in Figure 2. We implemented the GAN in python using Tensorflow and Keras. The architecture was modified from [2], however, changes were made in order to make it weakly supervised.

3.1 Architecture

The discriminator CNN was build using Keras sequential API. The discriminator contains five convolutional layers which were then flattened, a fully-connected layer with 1024 neurons, ending with a sigmoid activation function at the end. The first, second, and fifth convolutional layers of the discriminator are strided with a step size of 4, 2, and 2 respectively while the others are strided with a step size of 1. All of the layers use 'same' padding, meaning that the images will be surrounded with 'zeros'.

The generator is a fully convolutional residual CNN comprised of four residual blocks. We built the residual block in Keras using the sequential API. It contains two convolutional layers each followed with a ReLu activation function. The four blocks are fed into each other and are topped off with three more convolutional layers.

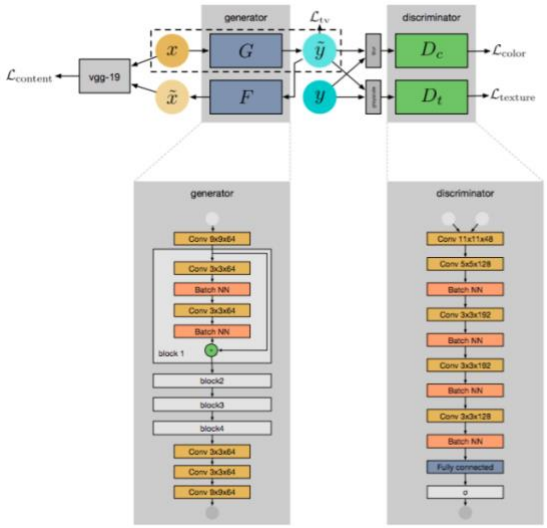

Figure 2: Proposed WESPE Architecture

To measure the consistency between the mapping and the input, we define the content consistency loss. Our loss functions consist of 2 adversarial discriminators (DC and DT respectively) and total variation. The purpose of DC and DT was to differentiate between the high-quality image and the enhanced image based on image color and texture respectively. Total Variation loss was to regularize towards the smoother results. Each of these loss functions is described below.

3.2 Content Consistency Loss

Content consistency loss is defined to measure the content similarity between the original image and the enhanced image. Content consistency aims to

preserve the contents of the image. It is defined on image input domain X i.e. on x and its reconstruction x'. The pixel level losses were restrictive, so we choose perceptual content loss based on Relu activations. It is defined as the l2 norm between features of the original image and the reconstructed image

$$L = \frac{1}{CjHjWj} ||vj(x) - vj(x')||$$

Where C_j, H_j, W_j represents number, height and width of the feature maps respectively.

3.3 Adversarial Color Loss

The quality of the image is measured by the adversarial discriminator that is trained to differentiate between the blurred images and high-quality images.

$$y_b(i,j) = \sum_{k,l} y(i+k, j+l) \cdot G_{k,l},$$

We want our discriminator to learn the differences contrast, brightness and various color loss between low quality and high quality images.

$$\mathcal{L}_{\text{color}} = -\sum_{i} \log D_c(G(x)_b).$$

Enhanced images have the similar color distributions as of the high-quality ones.

3.3 Adversarial Texture Loss

The role of the discriminator is to assess the image quality in terms of texture. The discriminator distinguishes between the enhanced image and high-quality images based on the texture. It is applied to the grayscale images and trained to check whether the given image was artificially enhanced or is a native high quality. The loss function is defined as:

$$\mathcal{L}_{\text{texture}} = -\sum_{i} \log D_t(G(x)_g).$$

Minimizing the texture loss will result in images of native high-quality ones.

3.4 TV Loss

The role of this loss is to impose spatial smoothness if the images. The loss is defined as:

$$\mathcal{L}_{\text{tv}} = \frac{1}{CHW} \|\nabla_x G(x) + \nabla_y G(x)\|,$$

where C,H,J are the dimensions of the generated image

3.5 Total Loss

The final WESPE loss is the summation of all the loss functions that are discussed above.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{content}} + 5 \cdot 10^{-3} (\mathcal{L}_{\text{color}} + \mathcal{L}_{\text{texture}}) + 10\, \mathcal{L}_{\text{tv}}.$$

4. Evaluation

We will discuss the evaluation methods and results of [1] as we believe that our GAN would closely follow the results of the original paper. The authors proposed several evaluation measures such as a full-reference evaluation using the DPED dataset which was used for supervised learning [2], no-reference evaluation using data from the wild, a user study, and finally the author's automated users Flickr [7] behavior. The experiments were conducted using several different cameras and datasets and compared against a commercial software baseline and Ignatov et al. The commercial software baseline that was used is Apple's Photo Enhancer (APE) which is standard on all Apple mobile devices.

4.1 Full-Reference Evaluation

The first test done was a full-reference evaluation using the DPED dataset. Since the DPED dataset was trained under supervised learning, we have pixel-aligned ground truth before and after data. The authors used this pixel-aligned data and used full-reference image quality metrics such as Point-Signal-to-Noise (PSNR) and Structural Similarity Index Measure (SSIM) to compare the enhanced test images to the ground truth DSLR quality photos.

PSNR compares the strength of the desired signal compared against the level of its background noise. SSIM meanwhile compares the similarity between the two photos. The results of the test showed that WESPE trained on the DIV2K dataset degrades the

SSIM and PSNR scores, but remained superior to APE. This was to be expected as we are measuring the enhanced images compared to a ground truth which is in the domain of DPED, not DIV2K.

4.2 No-reference evaluation

The next evaluation method was using a no-reference evaluation using images from the wild. Thanks to WESPE not requiring aligned image pairs, we are able to evaluate using new no-reference evaluation metrics and comparing using our WESPE method on images from several publicly available datasets such as Kitti and Cityscapes as described above as well as higher-end mobile cameras such as the iPhone 6. The metrics that are used for this evaluation are all no-reference quality metrics. These metrics will give an absolute image quality score instead of the previous method of getting a proximity to a reference photo.

These evaluation metrics are methods are the Codebook Representation for No-Reference Image Assessment (CORNIA) which is a "perceptual measure mapping to average human quality assessments for images" [6], and the common signal processing measures; Entropy (based on pixel level observations), and Bits Per Pixel (BPP). Both Entropy and BPP are indicators of the quantity of information in an image. As seen in Table 1 and Table 5, WESPE improves the original "wild" datasets of KITTI and Cityscapes by a large amount and even surpasses the APE in Table 4. However, we see that WESPE does improve the higher end cameras but not by a large amount. We conclude that WESPE does a greater job of healing images that are of poor quality than those of high-quality. This confirms our previous assumption that the similarity to the ground truth is not the matter of utmost importance.

	Original		
Images	Entropy	BPP	CORNIA
Cityscapes	6.73	8.44	43.42
KITTI	7.12	7.76	55.69
HTC One M9	7.51	9.52	**23.31**
Huawei P9	7.71	10.60	**20.63**
iPhone 6	7.56	11.65	**24.67**

Table 3: BPP, Entropy, and CORNIA on Original image on 5 "wild" datasets. Best results in bold.

	APE		
Images	Entropy	BPP	CORNIA
Cityscapes	7.30	6.74	46.73
KITTI	**7.58**	10.21	**37.64**
HTC One M9	7.64	9.62	28.46
Huawei P9	**7.78**	10.27	25.85
iPhone 6	**7.57**	9.25	35.82

Table 4: Average BPP, Entropy, and CORNIA on APE image on 5 "wild" datasets. Best results in bold.

	WESPE[DIV2K]		
Images	Entropy	BPP	CORNIA
Cityscapes	**7.56**	**11.59**	**32.53**
KITTI	7.55	**11.88**	39.09
HTC One M9	**7.69**	**12.99**	26.35
Huawei P9	7.70	**12.61**	27.52
iPhone 6	7.53	**13.44**	28.51

Table 5: Average BPP, Entropy, and CORNIA on WESPE[DIV2K] image on 5 "wild" datasets. Best results in bold.

4.3 User Study

The previous methods still do not tell us which photo a user actually prefers. Just because a photo has better CORNIA and BPP, doesn't mean that the image is more appealing to the user. The next experiment that was conducted was in order to find out which image users actually prefer. The study was conducted on the five "wild" datasets, comparing WESPE to both APE and the original photos. To measure the subjective quality of the image, 38 users were shown two images, either WESPE and Original, or WESPE and APE. The user had to choose which of the two they preferred. The results show that WESPE on average was the preferred image significantly outperforming APE and the original photo on the low-quality datasets (Cityscapes and KITTI). However, users found it more difficult to distinguish between the quality of WESPE photos when the images were already of high-quality.

Setting	**Cityscapes**	**Kitti**
WESPE vs ORIGINAL	0.94+0.03	0.81+0.10
WESPE vs APE	0.96+0.03	0.65+0.016

Table 4: User's preference on the two datasets using a pairwise comparison

4.4 Flickr Fave Scores

Getting uses to choose which picture they prefer is both slow and cumbersome. In order to mitigate this issue, [1] trained a method to mimic the user's behavior on the "World largest photographer focused community," Flickr [7].

The working assumption in create the Flickr Fav Score (FFS) method is that users add images that are of higher quality to their 'Faves'. The FFS score is the number of times an image is 'faved' over the number of times an image is viewed. FFS(I) = #F (I)/#V (I). A binary label is applied to those images to label the images as low-quality or high-quality depending on if the image is below or above the mean respectively. We see that the FFS score act in according to previous observations that WESPE trained on DIV2K vastly improves the users perceived quality of the image as seen in Table 5.

Images	Original	WESPE[DIV2K]
Cityscapes	0.4075	0.4339
KITTI	0.3792	0.5415
HTC One M9	0.5194	0.6193
Huawei P9	0.5322	0.5705
iPhone 6	0.5516	0.7412
Average	0.4780	0.5813

Table 5: Per dataset average FFS

Conclusion

In this paper, we implemented the work by [1] WESPE – Weakly supervised solution for image enhancement. The main advantage of WESPE is that it does not require before and after training pairs as in the previously proposed approaches. It is trained to map the low-quality images into high quality images without requiring any relation between them. For this, only 2 datasets with high quality images scrapped from the internet and the low-quality images taken by the low-end devices are required. For this, the authors of [1] proposed a transitive architecture of the GAN and loss functions designed for accurate image assessment. The work was validated on several publicly available datasets. The experiments demonstrated that WESPE produces comparable or surpassing the traditional enhancers while not requiring any form of supervision.

Reference

[1] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey and Luc Van Gool. "WESPE: Weakly Supervised Photo Enhancer for Digital Cameras", arXiv preprint arXiv:1709.01118, 2017

[2] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey and Luc Van Gool. "DSLR-Quality Photos on Mobile Devices with Deep Convolutional Networks", in IEEE International Conference on Computer Vision (ICCV), 2017

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[5] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, July 2017

[6] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 1098–1105. IEEE, 2012.

[7] "Flickr," Flickr, 16-Dec-2018. [Online]. Available: https://www.flickr.com/. [Accessed: 01-Dec-2018].