WESPE: Weakly Supervised Photo Enhancer for Digital Cameras

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

CVPR 2018

Presented by: Felix Singerman - 7970742 , Rishabh Kukreja - 300086824

December 4, 2018

Agenda

1. Introduction
2. Contributions
3. Proposed method
    - Content Consistency Loss
    - Adversarial Color Loss
    - Adversarial Texture Loss
    - TV Loss
    - Sum of Losses
    - Network Architecture
4. Experiments
5. Summary
6. Q&A

Introduction

• Low-end and compact mobile cameras demonstrate limited photo quality mainly due to space, hardware and budget constraints.
• In this work, we propose a deep learning solution that translates photos taken by cameras with limited capabilities into DSLR-quality photos automatically.

• Yet, one of the major bottlenecks of these solutions is the need for strong supervision using matched before/after training pairs of images. This requirement is often the source of a strong limitation of color/texture transfer and photo enhancement methods.

• We tackle this problem by introducing a weakly supervised photo enhancer (WESPE) - a novel image-to-image GAN-based architecture.

• The proposed model is trained by weakly supervised learning: unlike previous works, there is no need for strong supervision in the form of a large annotated dataset of aligned original/enhanced photo pairs

- The sole requirement is two distinct datasets:
  one from the source camera,
  one composed of arbitrary high-quality
- Hence, our solution is repeatable for any camera: collecting the data and training can be achieved in a couple of hours.
- The experiments demonstrate that WESPE produces comparable or improved qualitative results with state-of-the-art strongly supervised methods

Weakly Supervised Learning

Strongly Supervised Learning

Contributions

- Enhanced images improve the non enhanced ones in several aspects, including colorization, resolution and sharpness. Our contributions include:
- WESPE, a generic method for learning a model that enhances source images into DSLR-quality ones,
- A transitive CNN-GAN architecture, made suitable for the task of image enhancement and image domain transfer by combining state of the art losses with a content loss expressed on the input image,
- Large-scale experiments on several publicly available datasets with a variety of camera types, including subjective rating and comparison to the state of the art enhancement methods,
- A Flickr Faves Score (FFS) dataset consisting of 16K HD resolution Flickr photos with an associated number of likes and views that we use for training a separate scoring CNN to independently assess image quality of the photos throughout our experiments

Proposed Method

- Goal - To learn a mapping from
  Dom(X) → Dom(Y)
  (Low Quality) (High Quality)

- Model consist of 2 mappings:
  1.Generative mapping
  2.Inverse Generative mapping

Generative Mapping + Inverse Generative Mapping

$G : X \rightarrow Y$ $F : Y \rightarrow X.$

- To measure content consistency, we require content loss b/w input x and G(x) which is defined b/w original image and reconstructed image ẋ

  x = Original Image
  x = Reconstructed Image F(G(x))

Objective

1. Content Consistency Loss
    - To ensure G(x) preserve the x's content
2. Two Adversarial Losses
    - Color Loss
    - Texture Loss
3. Total Variation Loss
    - For smoother results

Content Consistency Loss

- We define the content consistency loss in the input image domain X i.e on x and x

x = Original Image $\tilde{x} = (F \circ G)(x)$,

- Our network is trained for both generative mapping G(x) as well as for inverse generative mapping F(X) and is aimed towards strong content similarity b/w original and enhanced images
- Pixel losses are restrictive, so we choose perceptual Content loss based on Relu activation

$$\mathcal{L}_{\text{content}} = \frac{1}{C_j H_j W_j} \|\psi_j(x) - \psi_j(\tilde{x})\|,$$

where ψ_j is the feature map from the j-th VGG-19 convolutional layer and C_j, H_j and W_j are the number, height and width of the feature maps, respectively.

Adversarial Color Loss

- To measure color quality Dc (adversarial discriminator) is trained to differentiate b/w blurred versions of enhanced images and high quality images
- Main idea is to make the discriminator learn the differences in
    1. Brightness
    2. Contrast
    3. Major color differences b/w low and high quality images
- We avoid texture and content comparison by defining a constant sigma very small which is used in Gaussian blur

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

Thus, color loss forces the enhanced images to have similar color distributions as the target high-quality pictures.

Adversarial Texture Loss

- Similarly to color, image texture quality is also assessed by an adversarial discriminator Dt

- It is applied to grayscale images and is trained to predict whether a given image was artificially enhanced (˜ yg) or is a "true" native high-quality image (yg).

- The network is trained to minimize the cross entropy loss function, the loss is defined as:

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

As a result, minimizing this loss will push the generator to produce images of the domain of native high-quality ones.

Total Variation (TV) Loss

- To impose spatial smoothness of the generated images we also add a total variation loss defined as follows:

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

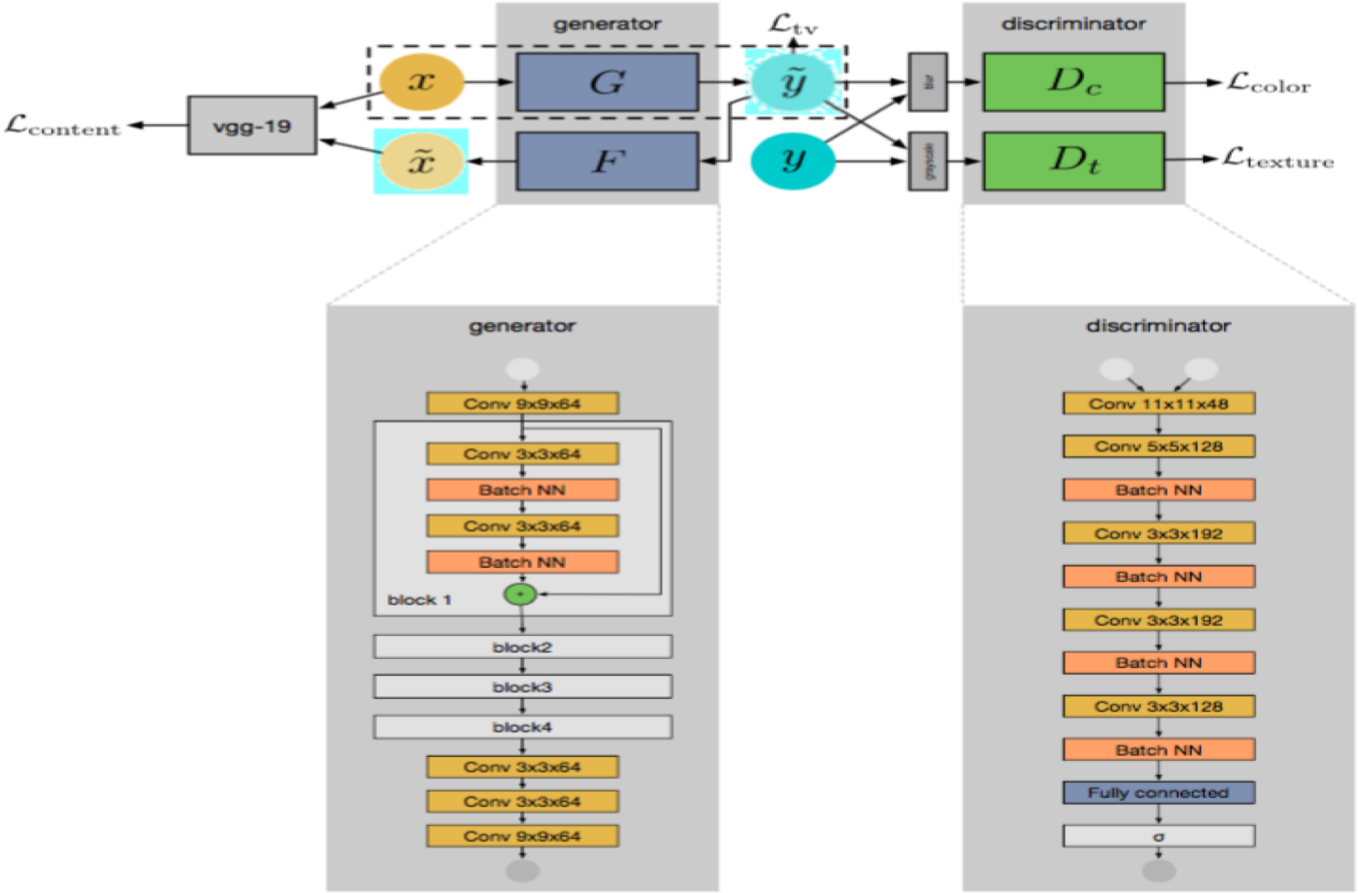
where C, H, W are dimensions of the generated image $G(x)$.

Sum of Losses

- The final WESPE loss is composed of a linear combination of the four mentioned losses:
    1. Content Loss
    2. Adversarial Color Loss
    3. Adversarial Texture Loss
    4. TV Loss

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

Network Architecture & Training Details



Experiments

Series of experiments covering several cameras and datasets

Compared against: Commercial software baseline, Ignatov et al.,

1. Full reference evaluation using DPED dataset used for supervised learning by Ignatov et al.
2. In the wild, as no ground truth needed
3. Subjective study using human raters
4. Flickr faves score emulator

Full-reference evaluation

Performed on the DPED dataset

Point Signal-to-Noise Ratio (PSNR)

Similarity index measure (SSIM)

Table 1: DPED dataset [13] with aligned images.

Camera source	Sensor	Image size	Photo quality	train images	test images
iPhone 3GS	3MP	2048 × 1536	Poor	5614	113
BlackBerry Passport	13MP	4160 × 3120	Mediocre	5902	113
Sony Xperia Z	13MP	2592 × 1944	Good	4427	76
Canon 70D DSLR	20MP	3648 × 2432	Excellent	5902	113

DPED images	APE		Weakly Supervised				Fully Supervised [13]	
			WESPE [DIV2K]		WESPE [DPED]			
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
iPhone	17.28	0.86	17.76	0.88	18.11	0.90	**21.35**	**0.92**
BlackBerry	18.91	0.89	16.71	0.91	16.78	0.91	**20.66**	**0.93**
Sony	19.45	0.92	20.05	0.89	20.29	0.93	**22.01**	**0.94**


iPhone


Apple Photo Enhancer


WESPE trained
on DPED


WESPE trained
on DIV2K


Strong
Supervised


DSLR

No-reference evaluation in the wild

No aligned image pairs from different cameras are available

Evaluate with no-reference quality metrics

CORNIA: perceptual measure mapping to average human quality assessments for images Signal Processing Measures: Quality of information in an image

- Entropy: Pixel level observation
- Bits Per Pixel (BPP)

Camera source	Sensor	Image size	Photo quality	train images	test images
KITTI [9]	N/A	1392 × 512	Poor	8458	124
Cityscapes [5]	N/A	2048 × 1024	Poor	2876	143
HTC One M9	20MP	5376 × 3752	Good	1443	57
Huawei P9	12MP	3968 × 2976	Good	1386	57
iPhone 6	8MP	3264 × 2448	Good	4011	57
Flickr Faves Score (FFS)	N/A	> 1600 × 1200	Poor-to-Excellent	15600	400
DIV2K [1]	N/A	~ 2040 × 1500	Excellent	900	0

No-reference evaluation in the wild

DPED images	Original			APE			[13]			WESPE [DPED]			WESPE [DIV2K]		
	entropy	bpp	CORNIA	entropy	bpp	CORNIA	entropy	bpp	CORNIA	entropy	bpp	CORNIA	entropy	bpp	CORNIA
iPhone	7.29	10.67	30.85	7.40	9.33	43.65	**7.55**	10.94	32.35	7.52	14.17	27.90	7.52	**15.13**	**27.40**
BlackBerry	7.51	12.00	11.09	7.55	10.19	23.19	7.51	11.39	20.62	7.43	12.64	23.93	**7.60**	**12.72**	**9.18**
Sony	7.51	11.63	32.69	**7.62**	11.37	34.85	7.53	10.90	**30.54**	7.59	12.05	34.77	7.46	**12.33**	34.56

Stronger improvement for low-quality cameras

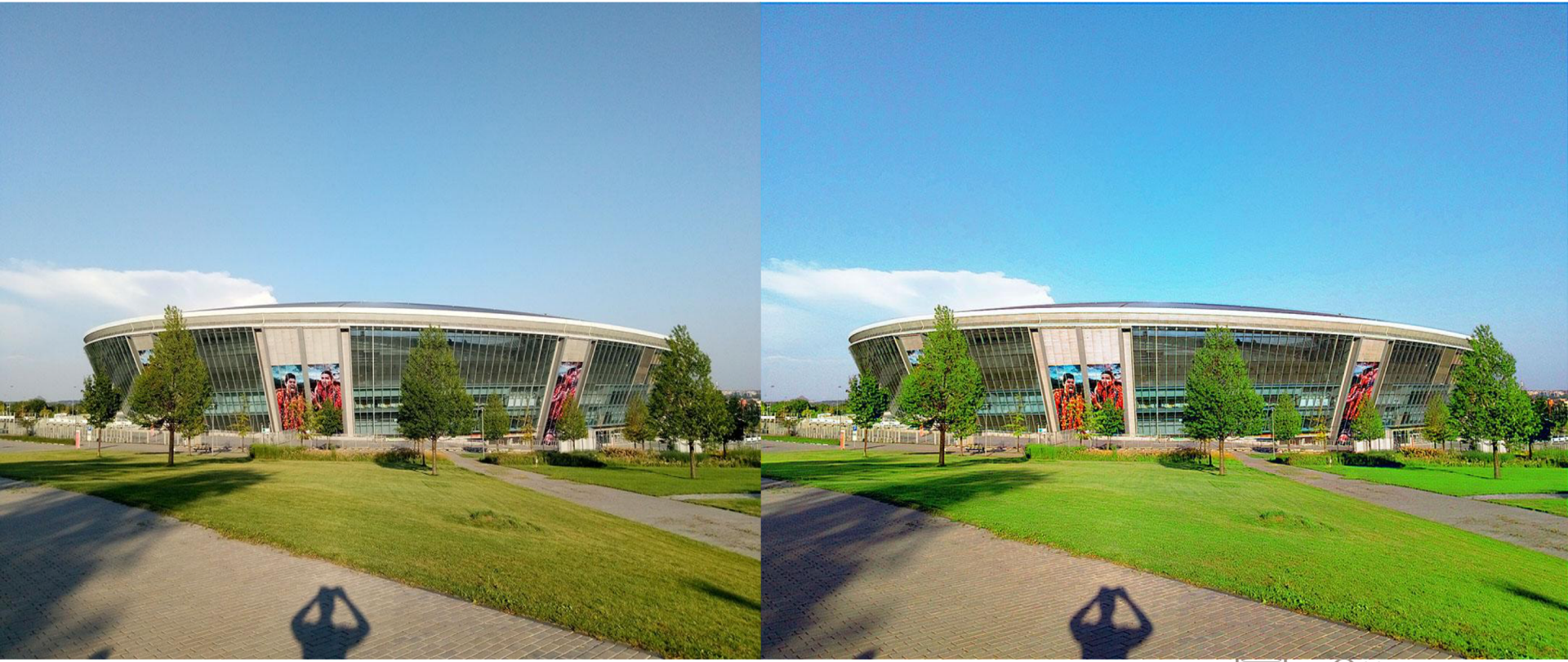
Proximity to ground truth is not the only matter of importance

Images	Original			APE			WESPE [DIV2K]		
	entropy	bpp	CORNIA	entropy	bpp	CORNIA	entropy	bpp	CORNIA
Cityscapes	6.73	8.44	43.42	7.30	6.74	46.73	**7.56**	**11.59**	**32.53**
KITTI	7.12	7.76	55.69	**7.58**	10.21	**37.64**	7.55	**11.88**	39.09
HTC One M9	7.51	9.52	**23.31**	7.64	9.64	28.46	**7.69**	**12.99**	26.35
Huawei P9	7.71	10.60	**20.63**	**7.78**	10.27	25.85	7.70	**12.61**	27.52
iPhone 6	7.56	11.65	**24.67**	**7.57**	9.25	35.82	7.53	**13.44**	28.51

User Study

- User study comparing subjective evaluation
- Original, APE-enhanced and WESPE-enhanced photos with DIV2K as target
- Pairwise forced choice method
- WESPE images are preferred on average

Setting	Cityscapes	KITTI	HTC M9	Huawei P9	iPhone 6
WESPE vs Original	0.94±0.03	0.81±0.10	0.73±0.08	0.63±0.11	0.70±0.10
WESPE vs APE	0.96±0.03	0.65±0.16	0.53±0.09	0.44±0.12	0.62±0.15



Flickr Faves Score

Virtual Rater to mimic Flickr user behavior

"World's largest photographer focused community" (Flickr.com)

Trained a binary classifier CNN to predict favourite status of an image by an average user

(FFS(I) = #F (I)/#V (I))

Trained a VGG19- style CNN to classify image Fave status

DPED images	original	fully Supervised [13]	Weakly Supervised WESPE [DPED] (ours)	WESPE [DIV2K] (ours)
iPhone	0.3190	0.5093	0.5341	**0.6155**
Blackberry	0.4765	0.5366	0.5904	**0.6001**
Sony	0.5694	0.6572	0.6774	**0.6828**
average	0.4550	0.5677	0.6006	**0.6328**

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**

Conclusion

- WESPE – a weakly supervised solution for the image quality enhancement problem
- Free of strong supervision in the form of aligned source-target training image pairs
- Map low-quality photos into the domain of high-quality photos without any correspondence between them
- Transitive architecture based on GANs and loss functions designed for accurate image quality assessment
- Comparative or superior results to past methods

References

A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool. WESPE: Weakly Supervised Photo Enhancer for Digital Cameras. *arXiv:1709.01118 [Cs]*. Retrieved from http://arxiv.org/abs/1709.01118, 2017

Thank you!!