This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

WESPE: Weakly Supervised Photo Enhancer for Digital Cameras

Andrey Ignatov[1], Nikolay Kobyshev[1], Radu Timofte[1], Kenneth Vanhoey[1], Luc Van Gool[1,2]

[1] Computer Vision Laboratory, ETH Zürich, Switzerland [2] ESAT - PSI, KU Leuven, Belgium

{andrey,nk,timofter,vanhoey,vangool}@vision.ee.ethz.ch

Abstract

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. We tackle this problem by introducing a weakly supervised photo enhancer (WESPE) – a novel image-to-image Generative Adversarial Network-based architecture. The proposed model is trained by under weak supervision: 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, and one composed of arbitrary high-quality images that can be generally crawled from the Internet – the visual content they exhibit may be unrelated. In this work, we emphasize on extensive evaluation of obtained results. Besides standard objective metrics and subjective user study, we train a virtual rater in the form of a separate CNN that mimics human raters on Flickr data and use this network to get reference scores for both original and enhanced photos. Our experiments on the DPED, KITTI and Cityscapes datasets as well as pictures from several generations of smartphones demonstrate that WESPE produces comparable or improved qualitative results with state-of-the-art strongly supervised methods.

1. Introduction

The ever-increasing quality of camera sensors allows us to photograph scenes with unprecedented detail and color. But as one gets used to better quality standards, photos captured just a few years ago with older hardware look dull and outdated. Analogously, despite incredible advancement in quality of images captured by mobile devices, compact sensors and lenses make DSLR-quality unattainable for them, leaving casual users with a constant dilemma of relying on their lightweight mobile device or transporting a heavier-weight camera around on a daily basis. However, the second option may not even be possible for a number of other applications such as autonomous driving or video surveillance systems, where primitive cameras are usually employed.

Figure 1: Cityscapes image enhanced by our method.

In general, image enhancement can be done manually (*e.g.*, by a graphical artist) or semi-automatically using specialized software capable of histogram equalization, photo sharpening, contrast adjustment, etc. The quality of the result in this case significantly depends on user skills and allocated time, and thus is not doable by non-graphical experts on a daily basis, or not applicable in case of real-time or large-scale data processing. A fundamentally different option is to train various learning-based methods that allow to automatically transform image style or to perform image enhancement. 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 [23] and photo enhancement [13] methods.

In this paper, we present a novel weakly supervised solution for the image enhancement problem to deliver ourselves from the above constraints. That is, we propose a deep learning architecture that can be trained to enhance im-

ages by mapping them from the domain of a given source camera into the domain of high-quality photos (supposedly taken by high-end DSLRs) while not requiring any correspondence or relation between the images from these domains: only two separate photo collections representing these domains are needed for training the network. To achieve this, we take advantage of two novel advancements in Generative Adversarial Networks (GAN) and Convolutional Neural Networks (CNN): **i)** transitive CNNs to map the enhanced image back to the space of source images so as to relax the need of paired ground truth photos [36], and **ii)** loss functions combining color, content and texture loss to learn photorealistic image quality [13]. The key advantage of the method is that it can be learned easily: the training data is trivial to obtain for any camera and training takes just a few hours. Yet, quality-wise, our results still surpass traditional enhancers and compete with state of the art (fully supervised) methods by producing artifact-less results.

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,
- openly available models and code[1], that we progressively augment with additional camera models / types.

2. Related work

Automatic photo enhancement can be considered as a typical – if not the ultimate – computational photography task. To devise our solution, we build upon three sub-fields: style transfer, image restoration and general-purpose image-to-image enhancers.

2.1. Style transfer

The goal of style transfer is to apply the style of one image to the (visual) content of another. Traditional texture/color/style transfer techniques [7, 11, 20, 23] rely on an exemplar before/after pair that defines the transfer to be applied. The exemplar pair should contain visual content having a sufficient level of analogy to the target image's content which is hard to find, and this hinders its automatic and mass usage. More recently, neural style transfer alleviates this requirement [8, 29]. It builds on the assumption that the shallower layers of a deep CNN classifier – or more precisely, their correlations – characterize the style of an image, while the deeper ones represent semantic content. A neural network is then used to obtain an image matching the style of one input and the content of another. Finally, generative adversarial networks (GAN) append a discriminator CNN to a generator network [10]. The role of the former is to distinguish between two domains of images: *e.g.*, those having the style of the target image and those produced by the generator. It is jointly trained with the generator, whose role is in turn to fool the discriminator by generating an image in the right domain, *i.e.*, the domain of images of correct style. We exploit this logic to force the produced images to be in the domain of target high-quality photos.

2.2. Image restoration

Image quality enhancement has traditionally been addressed through a list of its sub-tasks, like super-resolution, deblurring, dehazing, denoising, colorization and image adjustment. Our goal of hallucinating high-end images from low-end ones encompasses all these enhancements. Many of these tasks have recently seen the arrival of successful methods driven by deep learning phrased as image-to-image translation problems. However, a common property of these works is that they are targeted at *removing artifacts added artificially* to clean images, thus requiring to model all possible distortions. Reproducing the flaws of the optics of one camera compared to a high-end reference one is close to impossible, let alone repeating this for a large list of camera pairs. Nevertheless, many useful ideas have emerged in these works, their brief review is given below.

The goal of *image super-resolution* is to restore the original image from its downscaled version. Many end-to-end CNN-based solutions exist now [6, 16, 22, 25, 28]. Initial methods used pixel-wise mean-squared-error (MSE) loss functions, which often generated blurry results. Losses based on the activations of (a number of) VGG-layers [15] and GANs [17] are more capable of recovering photorealistic results, including high-frequency components, hence produce state of the art results. In our work, we incorporate both the GAN architectures and VGG-based loss functions.

Image colorization [4, 21, 34], which attempts to regress the 3 RGB channels from images that were reduced to single-channel grayscale, strongly benefits from the GAN architecture too [14]. Image *denoising*, *deblurring* and *dehazing* [3, 12, 19, 27, 35], photographic *style control* [31] and *transfer* [18], as well as *exposure correction* [33] are

[1] http://people.ee.ethz.ch/~ihnatova/wespe.html

another improvements and adjustments that are included in our learned model. As opposed to mentioned related work, there is no need to manually model these effects in our case.

2.3. General-purpose image-to-image enhancers

We build our solution upon very recent advances in image-to-image translation networks. Isola *et al*. [14] present a general-purpose translator that takes advantage of GANs to learn the loss function depending on the domain the target image should be in. While it achieves promising results when transferring between very different domains (*e.g*., aerial image to street map), it lacks photorealism when generating photos: results are often blurry and with strong checkerboard artifacts. Compared to our work, it needs *strong supervision*, in the form of many before/after examples provided at training time.

Zhu *et al*. [36] loosen this constraint by expressing the loss in the space of input rather than output images, taking advantage of a backward mapping CNN that transforms the output back into the space of input images. We apply a similar idea in this work. However, our CNN architecture and loss functions are based on different ideas: fully convolutional networks and elaborated losses allow us to achieve photorealistic results, while eliminating typical artifacts (like blur and checkerboard) and limitations of encoder-decoder networks.

Finally, Ignatov *et al*. [13] propose an end-to-end enhancer achieving photorealistic results for arbitrary-sized images due to a composition of content, texture and color losses. However, it is trained with a strong supervision requirement for which a dataset of aligned ground truth image pairs taken by different cameras was assembled (*i.e*., the DPED dataset). We build upon their loss functions to achieve photorealism as well, while adapting them to the new architecture suitable for our weakly supervised learning setting. While we do not need a ground truth aligned dataset, we use DPED to report the performance on. Additionally, we provide the results on public datasets (KITTI, Cityscapes) and several newly collected datasets for smartphone cameras.

3. Proposed method

Our goal is to learn a mapping from a source domain X (*e.g*., defined by a low-end digital camera) to a target domain Y (*e.g*., defined by a collection of captured or crawled high-quality images). The inputs are unpaired training image samples $x \in X$ and $y \in Y$. As illustrated in Figure 2, our model consists of a generative mapping $G : X \rightarrow Y$ paired with an inverse generative mapping $F : Y \rightarrow X$. To measure content consistency between the mapping $G(x)$ and the input image x, a content loss based on VGG-19 features is defined between the original and reconstructed images x and $\tilde{x} = (F \circ G)(x)$, respectively. Defining the *content loss* in

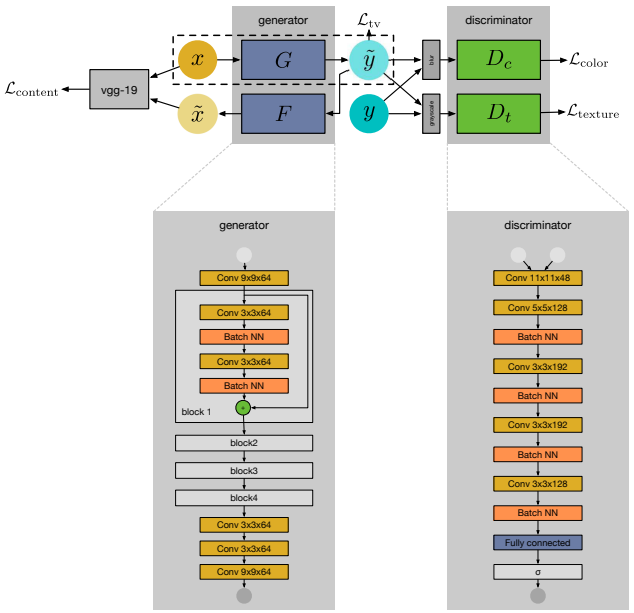

Figure 2: Proposed WESPE architecture.

the input image domain allows us to circumvent the need of before/after training pairs. Two adversarial discriminators D_c and D_t and total variation (TV) complete our loss definition. D_c aims to distinguish between high-quality image y and enhanced image $\tilde{y} = G(x)$ based on image colors, and D_t based on image texture. As a result, our objective comprises: i) content consistency loss to ensure G preserves x's content, ii) two adversarial losses ensuring generated images $\tilde{y}$ lie in the target domain Y: a color loss and a texture loss, and iii) TV loss to regularize towards smoother results. We now detail each of these loss terms.

3.1. Content consistency loss. We define the *content consistency loss* in the input image domain X: that is, on x and its reconstruction $\tilde{x} = F(\tilde{y}) = F \circ G(x)$ (inverse mapping from the enhanced image), as shown in Figure 2. Our network is trained for both the direct G and inverse F mapping simultaneously, aiming at strong content similarity between the original and enhanced image. We found pixel-level losses too restrictive in this case, hence we choose a perceptual content loss based on ReLu activations of the VGG-19 network [26], inspired by [13, 15, 17]. It is defined as the l_2-norm between feature representations of the input image x and the recovered image $\tilde{x}$:

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

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.

3.2. Adversarial color loss. Image color quality is measured using an adversarial discriminator D_c that is trained

to differentiate between the blurred versions of enhanced $\tilde{y}_b$ and high-quality y_b images:

$$y_b(i,j) = \sum_{k,l} y(i+k, j+l) \cdot G_{k,l}, \tag{2}$$

where $G_{k,l} = A \exp\left(-\frac{(k-\mu_x)^2}{2\sigma_x} - \frac{(l-\mu_y)^2}{2\sigma_y}\right)$ defines Gaussian blur with $A = 0.053$, $\mu_{x,y} = 0$, and $\sigma_{x,y} = 3$ set empirically.

The main idea here is that the discriminator should learn the differences in brightness, contrast and major colors between low– and high-quality images, while it should avoid texture and content comparison. A constant σ was defined experimentally to be the smallest value that ensures texture and content eliminations. The loss itself is defined as a standard generator objective, as used in GAN training:

$$\mathcal{L}_{\text{color}} = -\sum_i \log D_c(G(x)_b). \tag{3}$$

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

3.3. Adversarial texture loss. Similarly to color, image texture quality is also assessed by an adversarial discriminator D_t. This is applied to grayscale images and is trained to predict whether a given image was artificially enhanced ($\tilde{y}_g$) or is a "true" native high-quality image (y_g). As in the previous case, the network is trained to minimize the cross-entropy loss function, the loss is defined as:

$$\mathcal{L}_{\text{texture}} = -\sum_i \log D_t(G(x)_g). \tag{4}$$

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

3.4. TV loss. To impose spatial smoothness of the generated images we also add a total variation loss [2] defined as follows:

$$\mathcal{L}_{\text{tv}} = \frac{1}{CHW} \|\nabla_x G(x) + \nabla_y G(x)\|, \tag{5}$$

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

3.5. Sum of losses. The final WESPE loss is composed of a linear combination of the four aforementioned losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{content}} + 5 \cdot 10^{-3} \left(\mathcal{L}_{\text{color}} + \mathcal{L}_{\text{texture}}\right) + 10\, \mathcal{L}_{\text{tv}}. \tag{6}$$

The weights were picked based on preliminary experiments on our training data.

3.6. Network architecture and training details. The overall architecture of the system is illustrated in Figure 2. Both generative and inverse generative networks G and F are fully-convolutional residual CNNs with four residual blocks, their architecture was adapted from [13]. The discriminator CNNs consist of five convolutional and one fully-connected layer with 1024 neurons, followed by the last layer with a sigmoid activation function on top of it. The first, second and fifth convolutional layers are strided with a step size of 4, 2 and 2, respectively. For each dataset the train/test splits are as shown in Tables 2 and 4.

The network was trained on an *NVIDIA Titan X* GPU for 20K iterations using a batch size of 30 and the size of the input patches was 100×100 pixels. The parameters of the networks were optimized using the Adam algorithm. The experimental setup was identical in all experiments.

4. Experiments

To assess the abilities and quality of the proposed network (WESPE), we apply a series of experiments covering several cameras and datasets. We also compare against a commercial software baseline (the Apple Photos image enhancement software, or APE, version 2.0) and the latest state of the art in the field by Ignatov *et al.* [13], that uses learning under full supervision. We start our experiments by doing a full-reference quantitative evaluation of the proposed approach in section 4.1, using the ground truth DPED dataset used for supervised training by Ignatov *et al.* [13]. WESPE however is unsupervised, so it can be applied to any dataset in the wild as no ground truth enhanced image is needed for training. In section 4.2 we apply WESPE on such datasets of various nature and visual quality, and evaluate quantitatively using no-reference quality metrics. Since the main goal of WESPE is qualitative performance which is not always reflected by conventional metrics, we additionally use subjective evaluation of the obtained results. Section 4.3 presents a study involving human raters, and in section 4.4 we build and use a Flickr faves score emulator to emulate human rating on a large scale. For all experiments, we also provide qualitative visual results.

4.1. Full-reference evaluation

In this section, we perform our experiments on the the DPED dataset (see Table 2) that was initially proposed for learning a photo enhancer with full supervision [13]. DPED is composed of images from three smartphones with low –to middle-end cameras (*i.e.*, iPhone 3GS, BlackBerry Passport and Sony Xperia Z) paired with images of the same scenes taken by a high-end DSLR camera (*i.e.*, Canon 70D) with pixel alignment. Thanks to this pixel-aligned ground truth before/after data, we can exploit full-reference image quality metrics to compare the enhanced test images with the ground truth high-quality ones. For this we use both the Point Signal-to-Noise Ratio (*PSNR*) and the structural similarity index measure (*SSIM*) [30]. The former measures the amount of signal lost w.r.t. a reference signal (*e.g.*, an image), the latter compares two images' similarity in terms of visually structured elements and is known for its improved correlation with human perception, surpassing PSNR.

Figure 3: From left to right, top to bottom: original iPhone 3GS photo and the same image after applying, resp.: Apple Photo Enhancer, WESPE trained on DPED, WESPE trained on DIV2K, Ignatov *et al*. [13], and the corresponding DSLR image.

We adhere to the setup of [13] and train our model to map source photos to the domain of target DSLR images for each of three mobile cameras from the DPED dataset separately. Note that we use the DSLR photos in weak supervision only (without exploiting the pairwise correspondence between the source/target images): the adversarial discriminators are trained at each iteration with a random positive (*i.e*., DSLR) image and a random negative (*i.e*., non-DSLR) one. For each mobile phone camera, we train two networks with different target images: first using the original DPED DSLR photos as target (noted "WESPE [DPED]"), second using the high-quality pictures from the DIV2K dataset [1] (noted WESPE [DIV2K]). Full-reference (PSNR, SSIM) scores calculated w.r.t. the DPED ground truth enhanced images are given in Table 1.

Our WESPE method trained with the DPED DSLR target performs better than the baseline method (APE). Considering the better SSIM metric only, it is even almost as good as the network in [13] that uses a fully supervised approach and requires pixel-aligned ground truth images. WESPE trained on DIV2K images as target (WESPE [DIV2K]) and tested w.r.t. DPED images degrades PSNR and SSIM scores compared to WESPE [DPED], but still remains above APE. This is unsurprising as we are measuring proximity to known ground truth images laying in the domain of DPED DSLR photos (and not DIV2K): being close to it does not necessarily imply looking good. Visually (see Figs. 3 and 4), WESPE [DIV2K] seem to show crisper colors and we hypothesize they may be preferable, albeit further away from the ground truth image. This also hints that using diverse data (DIV2K has a diverse set of sources) of high-quality images (*e.g*., with few noise) may be beneficial as well. The following experiments try to confirm this.

Table 1: Average PSNR and SSIM results on DPED test images. Best results are in **bold**.

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**

4.2. No-reference evaluation in the wild

WESPE does not require before/after ground truth correspondences to be trained, so in this section we train it on various datasets in the wild whose main characteristics are shown in Table 4 as used in our experiments. Besides computing no-reference scores for the results obtained in the previous section, we complement the DPED dataset containing photos from older phones with pictures taken by phones marketed as having state-of-the-art cameras: the iPhone 6, HTC One M9 and Huawei P9. To avoid compression artifacts which may occur in online-crawled images, we did a manual collection in a peri-urban environment of thousands of pictures for each phone/camera. We additionally consider two widely-used datasets in Computer Vision and learning: the Cityscapes [5] and KITTI [9] public datasets. They contain a large-scale set of urban images of low quality, which forms a good use case for automated

Table 2: 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

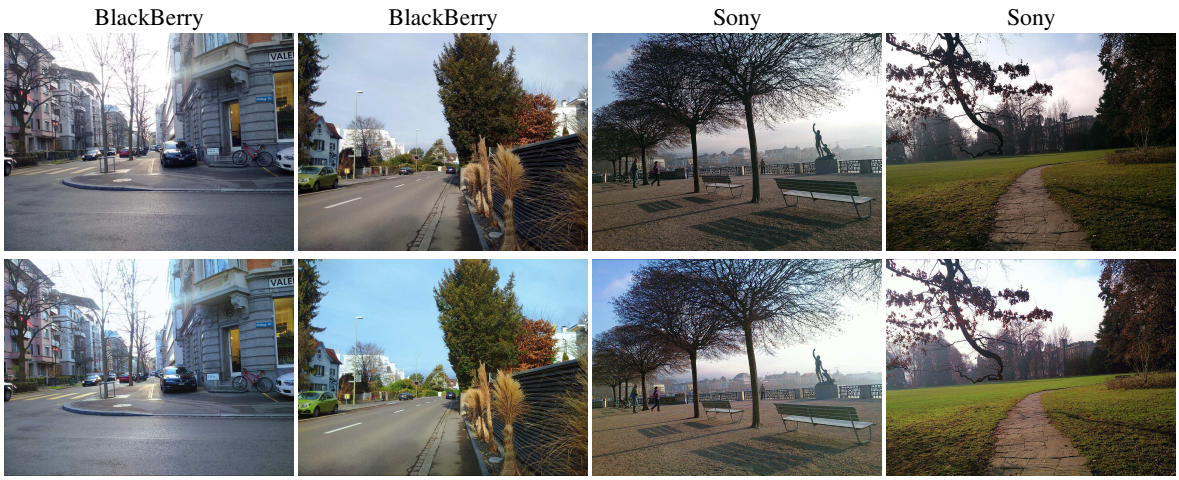

Figure 4: Original (top) vs. WESPE [DIV2K] enhanced (bottom) DPED images captured by BlackBerry and Sony cameras.

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	$> 1600 \times 1200$	Poor-to-Excellent	15600	400
DIV2K [1]	N/A	$\sim 2040 \times 1500$	Excellent	900	0

Table 4: Datasets in the wild as used in our experiments. No aligned image pairs from different cameras are available.

quality enhancement. That is, Cityscapes contains photos taken by a dash-camera (it lacks image details, resolution and brightness), while KITTI photos are brighter, but only half the resolution, disallowing sharp details (see Figure 5). Finally, we use the recent DIV2K dataset [1] of high quality images and diverse contents and camera sources as a target for our WESPE training.

Importantly, here we evaluate all images with no-reference quality metrics, that will give an absolute image quality score, not a proximity to a reference. For objective quality measurement, we mainly focus on the Codebook Representation for No-Reference Image Assessment (*CORNIA*) [32]: it is a perceptual measure mapping to average human quality assessments for images. Additionally, we compute typical signal processing measures, namely image *entropy* (based on pixel level observations) and bits per pixel (*bpp*) of the PNG lossless image compression. Both image entropy and bpp are indicators of the quantity of information in an image. We train WESPE to map from one of the datasets mentioned above to the DIV2K image dataset as target. We also report absolute quality measures (*i.e.*, bbp, entropy and CORNIA scores) on original DPED images as well as APE-enhanced, [13]-enhanced and WESPE-enhanced ([DPED] and [DIV2K] variants) images in Table 3, and take the best-performing methods to compare on the remaining datasets in Table 6.

Table 3 shows that the DIV2K variant of WESPE generates the best overall image quality, surpassing [13] and the WESPE variant that targets DPED DSLR images. This confirms the impression that proximity to ground truth is not the only matter of importance. This table also shows that improvement is stronger for low-quality camera's (iPhone and Blackberry) than for the better Sony camera, which probably benefits less from the WESPE image healing. Moreover, targeting the DIV2K image quality domain seems to improve over the DPED DSLR domain: WESPE [DIV2K] generally improves or competes with WESPE [DPED] and even the fully supervised [13] network.

On datasets in the wild (Table 6), WESPE and APE improve the original images on all metrics on the urban images (KITTI and Cityscapes). WESPE demonstrates significantly better results on the CORNIA and bpp metrics, but also on image entropy. Recall that KITTI and Cityscapes consist of images of poor quality, and our method is successful in healing such pictures. On the smartphones, whose pictures are already of high quality, our method shows improved bpp and slightly worse CORNIA scores, while keeping image entropy on par. The latter findings are quite ambiguous, since visual results for the urban (Figure 5) and phone datasets (Figure 6) demonstrate that there is a significant image quality difference that is not fully reflected

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

Table 3: Average entropy, bit per pixel and CORNIA (lower is better) results on DPED test images. Best results are in **bold**.

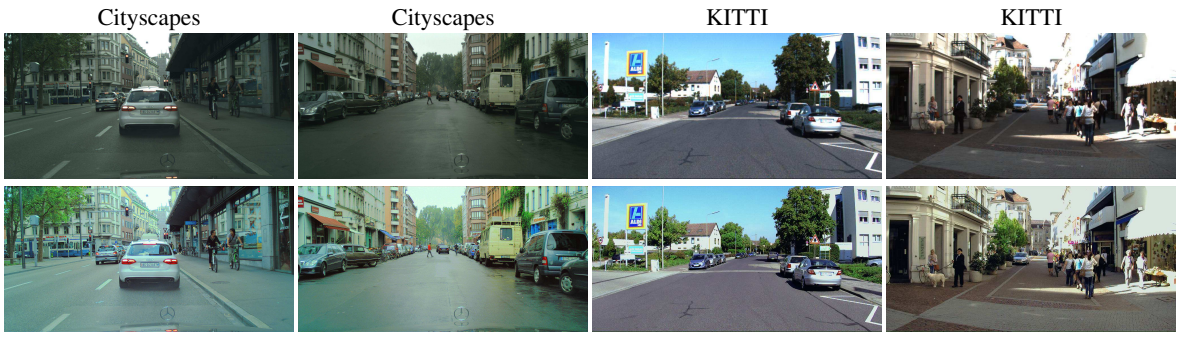

Figure 5: Examples of original (top) vs. enhanced (bottom) images for the Cityscapes and KITTI datasets.

Figure 6: Original (top) vs. enhanced (bottom) images for iPhone 6, HTC One M9 and Huawei P9 cameras.

by the entropy, bpp, and CORNIA quantitative numbers as proxy measures for perceived image quality. Moreover, since the correlation between objective scores and human perception can be debatable, in the following subsections we provide a complementary subjective quality evaluation.

4.3. User study

Since the final aim is to improve both the quality and aesthetics of an input image, we conducted a user study comparing subjective evaluation of the original, APE-enhanced and WESPE-enhanced photos with DIV2K as target, for the 5 datasets in the wild (see section 4.2 and Table 4). To assess subjective quality, we chose a pairwise forced choice method. The user's task was to choose the preferred picture among two displayed side by side. No additional selection criteria were specified, and users were allowed to zoom in and out at will without time restriction. Seven pictures were randomly taken from the test images of each dataset (*i.e.*, 35 pictures total). For each image, the users were shown a before vs. after WESPE-enhancement pair and a APE-enhanced vs. WESPE-enhanced pair to compare. 38 people participated in this survey and fulfilled the 35×2 selections. The question sequence, as well as the sequence of pictures in each pair were randomized for each user. Preference proportions for each choice are shown in Table 5.

WESPE-improved images are on average preferred over non-enhanced original images, even by a vast majority in the case of Cityscapes and KITTI datasets. On these two, the WESPE results are clearly preferred over the APE ones, especially on the Cityscapes dataset. On the modern phone cameras, users found it difficult to distinguish the quality of the WESPE-improved and APE-improved images, especially when the originals were already of good quality, on the HTC One M9 or Huawei P9 cameras in particular.

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

Table 5: User study results. The fraction of times WESPE result was preferred over original or APE-enhanced images.

Table 6: Average entropy, bit per pixel and CORNIA scores on five test datasets. Best results are in **bold**.

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

Table 7: FFS scores on the DPED dataset.

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**

Table 8: FFS scores on five test datasets in the wild.

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**

4.4. Flickr Faves Score

Gathering human-perceived photo quality scores is a tedious hence non-scalable process. To complement this, we train a virtual rater to mimic Flickr user behavior when adding an image to their favorites. Under the assumption that users tend to add better rather than lower quality images to their Faves, we train a binary classifier CNN to predict favorite status of an image by an average user, which we call the Flickr Faves Score (FFS).

First, we collect a *Flickr Faves Score dataset* (FFSD) consisting of 16K photos randomly crawled from Flickr along with their number of views and Faves. Only images of resolution higher than 1600×1200 pixels were considered and then cropped and resized to HD-resolution. We define the FFS score of an image as the number of times is was fav'ed over the number of times it was viewed ($FFS(I) = \#F(I)/\#V(I)$), and assume this strongly depends on overall image quality. We then binary-label all images as either low –or high-quality based the median FFS: below median is low-quality, above is high-quality. This naive methodology worked fine for our experiments (see results below): we leave analyzing and improving it for future work.

Next, we train a VGG19-style [26] CNN on random 224×224px patches to classify image Fave status and achieve 68.75% accuracy on test images. The network was initialized with VGG19 weights pre-trained on ImageNet [24], and trained until the early stopping criterion is met with a learning rate of 5e-5 and a batch size of 25. We split the data into training, validation and testing subsets of 15.2K, 400 and 400 images, respectively. Note that using HD-resolution inputs would be computationally infeasible while downscaling would remove image details and artifacts important for quality assessment. We used a single patch per image as more did not increase the performance.

We use this CNN to label both original and enhanced images from all datasets mentioned in this paper as Fave or not. In practice, we do this by averaging the results for five unique crops from each image (the identical crops are used for both original and enhanced photos). Per-dataset average FFS scores are shown in Tables 7 and 8. Note that this labeling differs from pairwise preference selection as in our user study of section 4.3: it's an absolute rating of images in the wild, as opposed to a limited pairwise comparison.

Our first observation is that the FFS scorer behaves coherently with all observations about DPED: the three smartphones' original photos that were termed as 'poor', 'mediocre' and 'average' in [13] have according FFS scores (Table 7, first column), and the more modern cameras have FFS scores that are similar to the best DPED smartphone (*i.e.*, Sony) camera (Table 8, first column). Finally, poorer-quality images in the Cityscapes and KITTI datasets score significantly lower. Having validated our scalable virtual FFS rater, one can note in Tables 7 and 8 that the FFS scores of WESPE consistently indicate improved quality over original images or the ones enhanced by the fully supervised method of [13]. Furthermore, this confirms our (now recurrent) finding that the [DIV2K] variant of WESPE improves over the [DPED] one.

5. Conclusion

In this work, we presented WESPE – a weakly supervised solution for the image quality enhancement problem. In contrast to previously proposed approaches that required strong supervision in the form of aligned source-target training image pairs, this method is free of this limitation. That is, it is trained to map low-quality photos into the domain of high-quality photos without requiring any correspondence between them: only two separate photo collections representing these domains are needed. To solve the problem, we proposed a transitive architecture that is based on GANs and loss functions designed for accurate image quality assessment. The method was validated on several publicly available datasets with different camera types. Our experiments reveal that WESPE demonstrates the performance comparable or surpassing the traditional enhancers and competes with the current state of the art supervised methods, while relaxing the need of supervision thus avoiding tedious creation of pixel-aligned datasets.

Acknowledgements

This work was partly supported by ETH Zurich General Fund (OK) and by Nvidia through a GPU grant.

References

[1] 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. 5, 6

[2] H. A. Aly and E. Dubois. Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing*, 14(10):1647–1659, Oct 2005. 4

[3] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, Nov 2016. 2

[4] Z. Cheng, Q. Yang, and B. Sheng. Deep colorization. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[5] 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. 5, 6

[6] C. Dong, C. C. Loy, K. He, and X. Tang. *Learning a Deep Convolutional Network for Image Super-Resolution*, pages 184–199. Springer International Publishing, Cham, 2014. 2

[7] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 341–346, New York, NY, USA, 2001. ACM. 2

[8] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015. 2

[9] 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, 6

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 2

[11] 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, pages 327–340, New York, NY, USA, 2001. ACM. 2

[12] M. Hradiš, J. Kotera, P. Zemčík, and F. Šroubek. Convolutional neural networks for direct text deblurring. In *Proceedings of BMVC 2015*. The British Machine Vision Association and Society for Pattern Recognition, 2015. 2

[13] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool. DSLR-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2, 3, 4, 5, 6, 8

[14] 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)*, July 2017. 2, 3

[15] J. Johnson, A. Alahi, and L. Fei-Fei. *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*, pages 694–711. Springer International Publishing, Cham, 2016. 2, 3

[16] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, June 2016. 2

[17] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016. 2, 3

[18] J.-Y. Lee, K. Sunkavalli, Z. Lin, X. Shen, and I. So Kweon. Automatic content-aware color and tone stylization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[19] Z. Ling, G. Fan, Y. Wang, and X. Lu. Learning deep transmission network for single image dehazing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2296–2300, Sept 2016. 2

[20] Y. Liu, M. Cohen, M. Uyttendaele, and S. Rusinkiewicz. Autostyle: Automatic style transfer from image collections to users' images. In *Computer Graphics Forum*, volume 33, pages 21–31. Wiley Online Library, 2014. 2

[21] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum. Natural image colorization. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 309–320. Eurographics Association, 2007. 2

[22] X. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2802–2810. Curran Associates, Inc., 2016. 2

[23] F. Okura, K. Vanhoey, A. Bousseau, A. A. Efros, and G. Drettakis. Unifying Color and Texture Transfer for Predictive Appearance Manipulation. *Computer Graphics Forum*, 2015. 1, 2

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 8

[25] 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)*, June 2016. 2

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 8

[27] P. Svoboda, M. Hradis, D. Barina, and P. Zemcík. Compression artifacts removal using convolutional neural networks. *CoRR*, abs/1605.00366, 2016. 2

[28] R. Timofte et al. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1110–1121, July 2017. 2

[29] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. *CoRR*, abs/1603.03417, 2016. 2

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. 4

[31] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu. Automatic photo adjustment using deep neural networks. *ACM Trans. Graph.*, 35(2):11:1–11:15, Feb. 2016. 2

[32] 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. 6

[33] L. Yuan and J. Sun. *Automatic Exposure Correction of Consumer Photographs*, pages 771–785. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. 2

[34] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. *ECCV*, 2016. 2

[35] X. Zhang and R. Wu. Fast depth image denoising and enhancement using a deep convolutional network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2499–2503, March 2016. 2

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 2, 3