计算机视觉课程报告模板

陈则衔
21821062
计算机科学与技术
zexianchen@zju.edu.cn

摘要

低端和紧凑型的移动相机出于空间，硬件和预算的限制使得照片质量有限。WESPE[1]提出了一种基于弱监督的深度学习的方法可以将质量较低的摄像机拍摄的照片自动转为DSLR质量的照片。该方法主要通过一个GAN神经网络构建而成。与之前的工作[2]在强监督下完成不同，该方法是虽然也需要两个不同的数据集。但这两个数据一个需要转变的低质量照片，一个数据集是高质量的DSLR照片，两者之间不需要存在一一对应关系。因此该方法可以适用于任何相机得到的照片。最后通过在DPED, KITTI, Cityscapes数据集和不同的智能手机拍摄的照片上进行实验，表明WESPE这一弱监督方法得到的记过和强监督方法得到的结果相比，具有相当的甚至更好的定性结果。

1. 论文简介

1.1. 论文背景

推荐的这篇论文WESPE出自于 2018 年CVPR的一篇文章，属于图像增强领域。该篇文章基于 2017 年ICCV上的文章《DSLR-Quality Photos on Mobile Devices With Deep Convolutional Networks》，并对其不足之处进行了改进。最重要的改进部分为将其强监督的方法改成了弱监督的方法。降低了对数据集的依赖性。因为现实中很难得到在不同分辨率质量下的相同场景的照片。值得一提的是，这两篇文章是由同一个实验室提出的:The Computer Vision Laboratory, ETH Zurich.并且这两篇文章均开源了所用的数据集和项目内容[3][4].

1.2. 相关工作

文章工作的主要目的在于，学习一个映射关系，可以从低质量的照片源映射到高质量的照片源。这和风格迁移、图像恢复以及通用的图像到图像增强器均有共通相似之处。

在图像风格迁移方面，是为了将一种图像风格应用到现有的图像上。现有的方法是通过生成对抗网络（GAN）和鉴别网络（CNN）相结合的办法，对抗生成结果。将生成网络生成的风格迁移结果的图像与真实的风格图传入鉴别网络达到目的。相机照片的图像增强也同样可以利用这种思路将生成图像映射到高质量照片空间。

图像增强在传统上可以分成多个不同的子问题，例如超分辨率，去模糊，去雾，去噪声，着色和图像调整。而从低质量图像加强到高质量图像覆盖了以上所有内容。该文章从中学习并利用了一些有用的方法。

例如图像超分辨率的目标是从缩小版本图像中恢复原始图像。原始的思路是直接用像素均方误差(pixel-wise mean-squared-error, MSE)来作为损失,但是其结果往往模糊。基于（多个）VGG层[5]和GAN [6]的激活的损失更能够恢复照片的真实结果，包括高频分量，从而产生最好的结果。该文章也利用到了VGG网络和GAN网络相结合的方法。

图像到图像转换网络与本文内容是最为接近的。本文通过这方面的最近进展构建了他们的解决方案。使用了与Zhu[7]等人相似的方法，通过输入空间中的损失，利用CNN后向映射将输出转回到输入图像的空间，来构造损失函数。同时，用与Ignatov[2]等人相同的损失函数来使得得到的照片与现实一致，并且将其从强监督的方式改成了弱监督的方法。

2. 论文特色

论文的特点在于将原来强监督的方法，变成了弱监督的学习方法。两者之间有着许多共同之处。我们先介绍强监督的方法，再介绍本文提出的弱监督方法，通过比较体现该文章的独特之处。

2.1. DPED Algorithm

该方法是一个端到端的方法，通过输入成对的低质量手机照片和高质量的DSLR质量的照片，从而学习到手机到DSLR相机的一种特征映射。其图像增强部分用 12 层残差卷积网络组成，该网络将手机照片作为输入

并且训练从而得到DSLR相机相对应的图像。整体网络结构如图 1 所示。

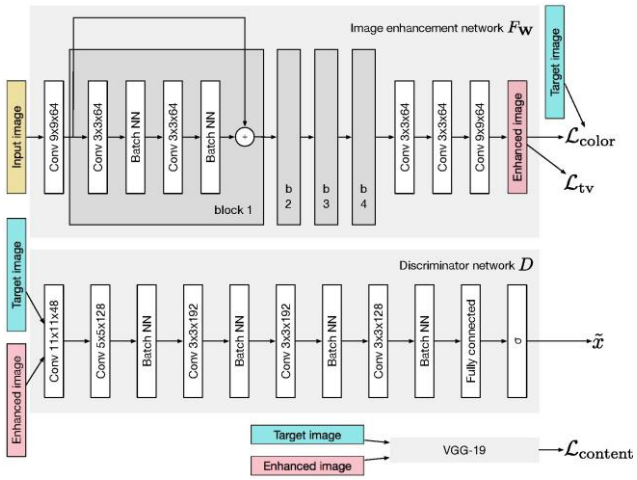

Figure 1 DEPD Model

可以发现整体的网络结构是一个基于GAN对抗网络的图像增强。图像增强部分是一个由 12 层残差卷积网络构成。判别网络由 5 层卷积层和一个全连接层构成，全连接层的神经元有 1024 个，最后通过sigmoid函数得到 2 维概率向量计算损失函数。除此之外，还有一个VGG-19 网络来进行content loss 的计算。综上该网络的损失函数由三个不同的部分组成。

2.1.1 Color loss：

增强的图像应该在颜色方面接近目标DSLR照片。为了测量他们之间的差异，先对两个图像应用高斯模糊，再对高斯模糊后的结果计算两者之间的欧式距离作为损失值。其具体内容如下：

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

其中X_b和Y_b分别是X和Y高斯模糊后的结果：

$X_b(i,j) = \sum_{k,l} X(i+k, j+l) \cdot G(k,l)$,

其中，2D高斯模糊操作如下:

$$G(k,l) = Aexp(-\frac{(k-u_x)^2}{2\sigma_x} - \frac{(l-u_y)^2}{2\sigma_y},$$

其中,$A = 0.053, u_{x,y} = 0, \sigma_{x,y} = 3$

2.1.2 Texture loss

为了测量增强图像的纹理质量，其训练了一个单独的对抗性CNN鉴别器，观察改进的和目标灰度图像，其目的是预测哪个图像是哪个。而图像增强网络的目的是欺骗这个CNN鉴别器，使其无法区分它们。

在对抗网络中，使用灰度图进行输入，原因在于图像的纹理信息与灰度空间分布有关。其具体损失值计算方式如下：

$$\mathcal{L}_{texture} = -\sum_i logD(F_W(I_s), I_t)$$

2.1.3 Content loss

分别将增强网络的图像和目标图像均输入到VGG-19网络中提取特征，原理在于如果增强网络学习到的图像和目标图像是一致的话，经过VGG-19 网络后提取到的特征也是接近的。其具体计算方式如下：

$$\mathcal{L}_{content} = \frac{1}{C_j H_j W_j} \| \varphi_j(F_W(I_s)) - \varphi_j(I_t) \|,$$

其中C_j, H_j, W_j代表了特征图的通道数量，高度和宽度，$F_W(I_s)$代表了图像增强的结果。

2.1.4 Total variation loss

梯度损失是整体上对图像进行微小的平滑，同时有效的去除椒盐噪声。其计算方式如下：

$$\mathcal{L}_{tv} = \frac{1}{CHW} \| \nabla_x F_W(I_s) + \nabla_y F_W(I_s) \|$$

其中，C, H, W是图像生成网络$F_W(I_s)$生成图像的各个维度值。

2.1.5 Total loss

最后整体的损失函数由前 4 个不同的损失函数构成，具体值如下所示：

$$\mathcal{L}_{total} = \mathcal{L}_{content} + 0.4 \cdot \mathcal{L}_{texture} + 0.1 \cdot \mathcal{L}_{color} + 400 \cdot \mathcal{L}_{tv}$$

2.2. WESPE Algorithm

WESPE的目标是学习一个从源数据域X（例如由低质量照片定义）到目标域Y（例如由高质量照片定义）的一个映射.输入的不成对训练图像为$x \in X, y \in Y$.整体模型如图 2 所示。可以发现，该模型的生成网络和鉴别网络和DPED模型是完全一致的。

模型的生成图像部分包括两个生成映射，分别是$G: X \to Y$和相反的生成映射，$F: Y \to X$.为了检测增强结果$G(x)$和输入图像x之间的内容一致性，同样用基于VGG-19 的特征图进行比较。比较对象分别为输入图像x和重构图像$\tilde{x} = (F \circ G)(x)$。定义输入图像域中的内容损失可以避免对训练对的需要。与DPED同样，还定义了颜色损失、纹理损失和total variation损失。不同之处还在于，这里用了 2 个对抗性鉴别器D_c, D_t分别对颜色和纹理信息进行鉴定判别。因此，整个模型目标如下:i)content loss保证G生成器能保留输入图像x的内容，ii)两个对抗性损失保证生成的图像$\tilde{y}$在目标域Y中：颜色损失和纹理损失,iii)通过TV损失来正则化，使得结果更为平滑。其各个损失具体内容如下:

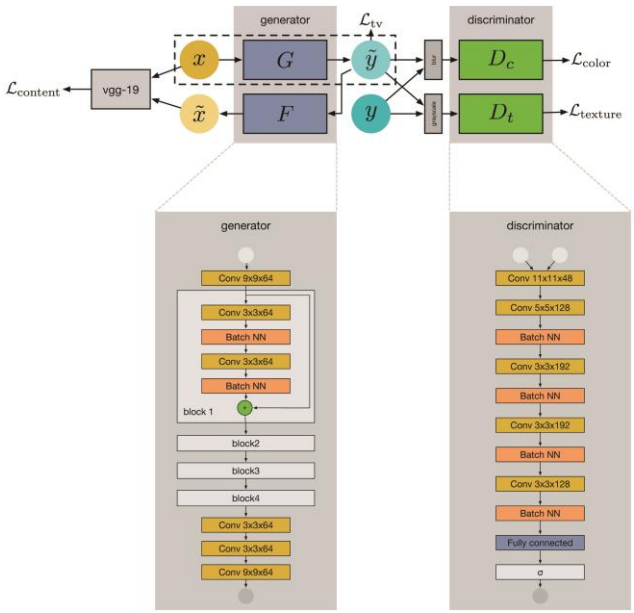

Figure 2 WESPE Model

2.2.1 Content consistency loss

为了保证增强图像的和原始图像内容的一致性而定义。计算方法与DPED相同，不同之处在比较的对象由强监督的训练对变成了输入图像和重构图像，具体如下：

$$\mathcal{L}_{content} = \frac{1}{C_j H_j W_j} \| \varphi_j(x) - \varphi_j(x) \|$$

其中，φ_j 代表了VGG-19 的第j层特征图。

2.2.2 Adversarial color loss

同样先增加高斯模糊，再对输入图像进行颜色损失的计算。这里采用交叉熵而不是MSE，由于不是强监督训练，因此直接采用MSE计算，会出现很大的损失偏差，用交叉熵更为合理，具体公式如下：

$$\mathcal{L}_{color} = -\sum_i logD_c(G(x)_b)$$

2.2.3 Adversarial texture loss

同样是对灰度图进行损失评估，与DPED一致，具体如下：

$$\mathcal{L}_{texture} = -\sum_i logD_t(G(x)_b)$$

2.2.4 TV loss

与DPED一致，具体如下:

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

2.2.5 Sum of losses

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

2.3. 总结

DEPD通过强监督下不同损失函数的设置，达到增强后的图片既能保持内容的一致性，又能使其拥有DSLR照片的特性。是一个比较典型的GAN神经网络。除此之外，内容损失借鉴了VGG网络和GAN网络的结合，通过特征层的损失值来进行评估，这往往能达到更好的效果。

WESPE和DEPD之间最重要的区别在于，WESPE将强监督转为了弱监督。因此其不能直接用相同的方法用MSE进行颜色损失的评估，从而建立了 2 个结构相同的CNN鉴别器。除此之外，要保持增强前后内容的一致性，缺少一一对应的标签图片，采用了 2 个生成网络，用重构图来代替标签图片，以保证内容的一致性。在内容一致性得到保证的情况下，将增强后图片的纹理和颜色属性映射到高质量图片域，从而完成增强。

3. 实验结果

论文从 4 个不同方面进行了实验评估。其中两个是定量评估，两个是定性评估。定量评估的对象分别是强监督DEPD方法和一个商业图片增强软件,APE(the Apple Photos image enhancement software, version2.0)分为有全面参考的评估，即有ground-truth，和没有ground-truth的数据集上进行评测。以定量评估为例给出实验结果。

3.1. 全面参考评估

其使用信噪比(Point Signal-to-Noise Ratio, PSNR)和结构相似性(structural similarity index measure, SSIM)[8]来判断增强图片的好坏。WESPE用两个不同的数据集，DPED和DIV2K[9]训练了 2 个不同的模型，在DEPD数据集上进行评测。实验数据为表 1，评测结果如表 2。

Table 1 DPED数据集内容

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

Table 2 实验结果，粗体代表最好的实验结果

DPED images	APE		Weakly Supervised				Fully Supervised	
			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**

从实验结果中可以发现，WESPE的方法基本要好于APE软件，但弱于强监督方法。其中，使用DEPED数据集的模型要在数值上好于使用DIV2K数据集的模型。这是好理解的，因为测试集也是DPED数据集，其映射域是同一个。观测图片结果如图 3 所示。

Figure 3 从左至右，从上至下为：原始iPhone照片，和相同的照片在APE,DPED数据集训练的WESPE，DIV2K数据训练的WESPE，强监督方法和对应的DSLR照片。

可以从图片中发现。在DIV2K数据集上训练的增强结果拥有清晰的颜色结果，尽管其训练数据和测试数据相距较远。这暗示着，用更离散的数据集进行训练可能得到更好的结果。下一个定量实验试图证明这一点。

3.2. 无参考评估的实验评估

在该实验中，使用了两个计算机视觉中通用的公共数据集，the Cityscapes[10]和KITTI[11].用无参考图像评估的数值表示(Codebook Representation for No-Reference Image Assessment, CORNIA)[12]指标以及图像熵(image entropy)和PNG无损图像压缩的每像位素(bits per pixel, bpp)来进行评估。评测的数据集和评测结果如表 3，表 4 所示。

Table 3 实验所用数据

Camera source	Sensor	Image size	Photo quality	train images	test images
KITTI	N/A	1392×512	Poor	8458	124
Cityscapes	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	N/A	$\sim 2040 \times 1500$	Excellent	900	0

Table 4 实验结果，粗体代表最好结果

DPED images	Original			APE			Fully Supervised			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

从实验结果中可以发现，基本上WESPE在DIV2K上的训练结果有着比较好的指标值。但有些图像，存在显着的图像质量差异，而熵，bpp和CORNIA定量数字并未完全反映出图像增强结果后的区别。因此该篇论文还通过了 2 个定性的实验从人的感知度上进一步证明本方法的优越性。篇幅所限不再赘述。

References

[1] Ignatov A, Kobyshev N, Timofte R, et al. WESPE: weakly supervised photo enhancer for digital cameras[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018: 691-700.

[2] Ignatov A, Kobyshev N, Timofte R, et al. DSLR-quality photos on mobile devices with deep convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3277-3285.

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

[4] http://people.ee.ethz.ch/~ihnatova/#title

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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. 4, 5, 6

[10] 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

[11] 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

[12] 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