UNDERWATER IMAGE ENHANCEMENT VIA LEARNING WATER TYPE DESENSITIZED REPRESENTATIONS

Zhenqi Fu, Xiaopeng Lin, Wu Wang, Yue Huang, Xinghao Ding ⋆

Key Laboratory of Underwater Acoustic Communication
and Marine Information Technology, Ministry of Education, Xiamen University
School of Informatics, Xiamen University, China
⋆dxh@xmu.edu.cn

ABSTRACT

We present a novel underwater image enhancement method termed SCNet to improve the image quality meanwhile cope with the degradation diversity caused by the water. SCNet is based on normalization schemes across both spatial and channel dimensions with the key idea of learning water type desensitized features. Specifically, we apply whitening to decorrelate activations across spatial dimensions for each instance in a mini-batch. We also eliminate channel-wise correlation by standardizing and re-injecting the first two moments of the activations across channels. The normalization schemes of spatial and channel dimensions are performed at each scale of the U-Net to obtain multi-scale representations. With such water type irrelevant encodings, the decoder can easily reconstruct the clean signal and be unaffected by the distortion types. Experimental results on two real-world underwater image datasets show that our approach can successfully enhance images with diverse water types, and achieves competitive performance in visual quality improvement.

***Index Terms*—** Underwater image, image enhancement, whitening, normalization, deep learning

1. INTRODUCTION

Underwater optical vision is a critical perception component for marine research and underwater robotics. For example, underwater surveillance systems and autonomous underwater vehicles rely on high-quality images to fulfill their objectives. Scientists also need clean underwater images to study deteriorating coral reefs and other aquatic life [1]. Unfortunately, the images acquired for these applications are commonly degraded due to various influences. One of the major factors is wavelength-dependent light absorption and scattering over the depth of objects in the scene. The absorption effect is caused by the fact that the red light is absorbed at a higher rate than green and blue in the water. Hence, underwater images are commonly dominated by bluish or greenish tint. The scattering phenomenon (including forward-scattering and backward-scattering) stems from suspending particles, which diminishes the image quality by introducing a homogeneous background noise and haze-like appearance.

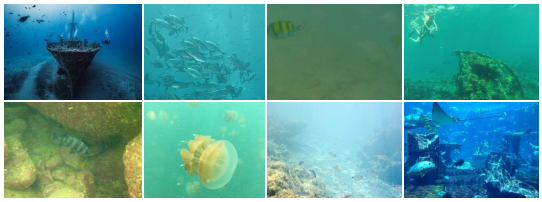


Fig. 1. Underwater scenes captured in diverse water types show a significant difference in appearances and styles.

Apart from the light attenuation, another challenge in underwater image enhancement (UIE) is the diversity of degradations. As presented in Fig. 1, underwater scenes captured in diverse water types (e.g., shallow coastal waters, deep oceanic waters, and muddy waters) show a significant difference in appearances and styles. Normally, it is difficult for a single model to enhance underwater images with such multiple degradation distributions. In other words, providing a universal solution for UIE is challenging.

Although numerous UIE approaches have been developed, such as model-free methods [2, 3], prior-based methods [4–8], and data-driven methods [9–12], few of them consider the challenges of degradation diversity explicitly. Anwar et al. [13] first synthesized ten different underwater image datasets, then they trained UIE models for each water type. However, this approach seems inefficient and relies on the prior knowledge of the water type for the given image. Since we do not know the water type ahead of time, Berman et al. [7] tried different parameter sets out of an existing library of water types (e.g., Jerlov water types). Each set leads to a different restored image and the one that best satisfies the Gray-World assumption is chosen as the final output. Simi-

The study is supported partly by the National Natural Science Foundation of China under Grants 82172033, U19B2031, 61971369, 52105126, China Postdoctoral Science Foundation (No.2021M702726), Science and Technology Key Project of Fujian Province(No.2019HZ020009) and Fundamental Research Funds for the Central Universities 20720200003.

larly, this approach is also inefficient because it must perform ten times for each image to be enhanced. Additionally, it is unreliable to select the best result by a simple Gray-World based quality assessment metric. Recently, Uplavikar et al. [14] learned domain agnostic features for multiple water types and generated clean versions from those features. Concretely, they additionally used a network to classify the water type of a given image from U-Net's [13] encodings. An adversarial loss is used to force the classifier to be unsure of the possible water types. Finally, water type agnostic representations can be learned by the U-Net's encoder.

In this paper, we propose SCNet, a normalization-based approach for UIE to improve the image quality meanwhile handle the diversity of water types. Generally, the degradation diversity of underwater images is introduced by both absorption and scattering. The former makes the image presents different colors. While the latter blurs the edges of objects and leads to various degrees of haze-like effects. Therefore, SCNet normalizes activations across both spatial and channel dimensions, which can reduce the impact of water types. As a result, SCNet can predict a cleaned image more accurately. In summary, this paper introduces the following contributions: 1) We present a normalization-based UIE method. Instead of designing a complex network architecture, we perform normalization schemes at each scale of a simple U-Net to learn multi-scale water type desensitized representations. 2) To obtain better enhancement performance, we not only de-correlate the activations across spatial dimensions but also normalize and re-inject the first two moments of the activation across channel dimensions. 3) Experimental results demonstrate that our approach outperforms the previous methods significantly in both improving the visual quality and dealing with the diversity of water types.

2. APPROACH

Our solution is based on normalization schemes, which standardize and whiten data using the extracted statistics. As shown in Fig. 2, we combine normalization schemes with the U-Net [13] to learn water type desensitized representations. Specifically, we perform spatial-wise and channel-wise normalization at each scale of the U-Net simultaneously.

2.1. Spatial-wise Normalization

Spatial-wise normalization is performed via instance whitening [16] to reduce the influence of diverse water types and discard the extracted statistics across spatial dimensions. We propose to adopt instance whitening to normalize features because the appearance of an individual image can be well encoded by the covariance matrix. In our method, spatial-wise normalization is performed in the U-net's skip-connection. Let $\mathbf{X} \in \mathbb{R}^{C\times NHW}$ refers to the data matrix of a mini-batch, where, C, N, H, W indicate the number of channels, number of instances, the height, and the width respectively. Here, N, H, and W are viewed as a single dimension for convenience. Let the matrix $\mathbf{X}_n \in \mathbb{R}^{C\times HW}$ be the n-th instance in the mini-batch, where $n \in \{1, 2, \ldots, N\}$. Then the whitening transformation Γ for an instance $\mathbf{X}_n$ can be formulated as:

$$\Gamma\left(\mathbf{X}_n\right) = \mathbf{\Sigma}^{-1/2}\left(\mathbf{X}_n - \boldsymbol{\mu}\right) \tag{1}$$

where $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ denote the mean vector and the covariance matrix computed from the data, respectively. Specifically, for instance whitening, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are calculated within each individual sample by:

$$\boldsymbol{\mu} = \frac{1}{HW}\mathbf{X}_n \tag{2}$$

$$\mathbf{\Sigma} = \frac{1}{HW}\left(\mathbf{X}_n - \boldsymbol{\mu}\right)\left(\mathbf{X}_n - \boldsymbol{\mu}\right)^T + \alpha\mathbf{I} \tag{3}$$

where α is a small positive number to prevent a singular $\mathbf{\Sigma}$. In this way, the whitening transformation Γ whitens each instance separately (i.e., $\Gamma\left(\mathbf{X}_n\right)\Gamma(\mathbf{X}_n)^T = \mathbf{I}$). Note that, in the covariance matrix $\mathbf{\Sigma}$, the diagonal elements are the variance for each channel, while the off-diagonal elements are the correlation between channels. Therefore, Eq. (1) cannot only standardize but also de-correlate the activations. To enhance the representation capacity, we add scale and shift operations for instance whitening. Thus, Eq. 1 can be rewritten as:

$$\Gamma\left(\mathbf{X}_n\right) = \mathbf{\Sigma}^{-1/2}\left(\mathbf{X}_n - \boldsymbol{\mu}\right)\gamma + \beta \tag{4}$$

where γ and β are learnable parameters denoting the scale and shift operations, respectively.

2.2. Channel-wise Normalization

Diverse water types lead to different degrees of scattering effects, which blur the image edge and reduce the visibility of important objects. Considering that channel-wise statistics are position-dependent and can well reveal the structural information about the input image and extracted features [17], we propose to leverage channel-wise normalization to further reduce the impact of degradation diversity. Concretely, we first remove the mean and the standard deviation across channels in U-Net's encoder. Naturally, the remaining representations are structure irrelevant. Although removing the first two moments does benefit training, it also eliminates important information about the image content, which would have to be painfully relearned in the decoder. Therefore, similar to [17], we re-inject them into the decoder. Specially, we use 1×1 convolutional operator to generate optimized statistics (multi-channel outputs). Similar to the notation definition in the spatial-wise normalization, let matrix $\mathbf{X}_n \in \mathbb{R}^{HW\times C}$ be the n-th sample in the mini-batch, where $n \in \{1, 2, \ldots, N\}$. Then the channel-wise normalization Ω for a sample $\mathbf{X}_n$ can be calculated as:

$$\Omega\left(\mathbf{X}_n\right) = \frac{\mathbf{X}_n - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{5}$$

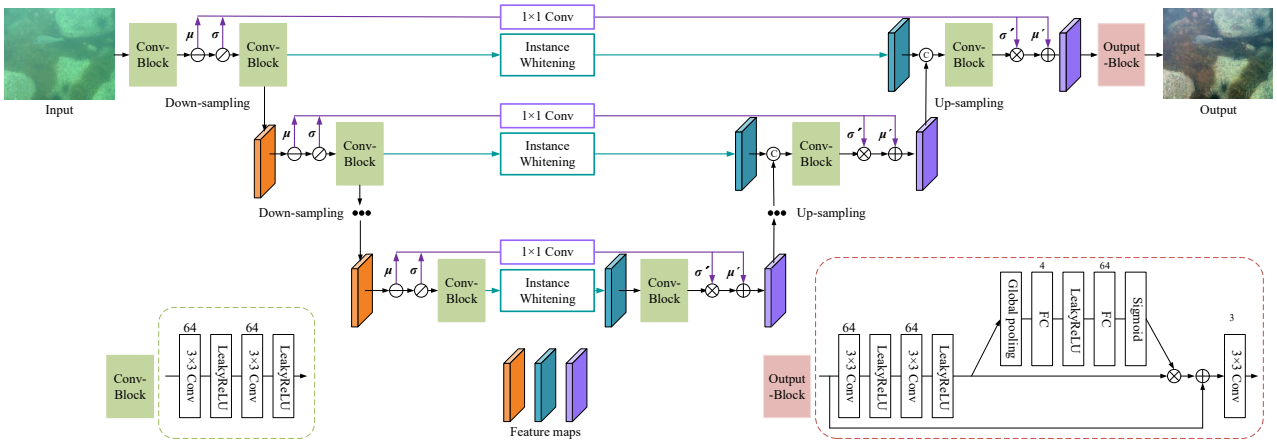


Fig. 2. Illustration of normalization across spatial and channel dimensions at each scale of the U-Net. The normalized activations are style and appearance irrelevant. The decoder can easily reconstruct the clean signal. We embed a SE [15] unit in the output block to improve the network modeling capacity.

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the mean and standard deviation vectors, respectively. $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are calculated by:

$$\boldsymbol{\mu} = \frac{1}{C}\mathbf{X}_n \tag{6}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{1}{C}\sum\left(\mathbf{X}_n - \boldsymbol{\mu}\right)^2 + \alpha\mathbf{I}} \tag{7}$$

where α is a small positive number to prevent a singular $\boldsymbol{\sigma}$. As mentioned before, after removing the mean and standard deviation across channel dimensions, we transform and re-inject them into the corresponding decoder layer. To be specific, the mean is added to the features, and the standard deviation is multiplied, which can be written as:

$$\mathbf{y} = \boldsymbol{\mu}'\mathbf{x} + \boldsymbol{\sigma}' \tag{8}$$

where $\boldsymbol{\mu}'$ and $\boldsymbol{\sigma}'$ denote the optimized statistics of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, respectively.

2.3. Loss Function

Given a training set $\{\mathbf{u}_{raw}, \mathbf{u}_{gt}\}$, where $\mathbf{u}_{raw}$ indicates the raw underwater instances and $\mathbf{u}_{gt}$ refers to the corresponding clean versions. We adopt mean squared error (MSE) and perceptual similarity (PS) [18] for training our network. MSE is calculated based on pixel-wise difference:

$$\ell_{\text{MSE}} = \frac{1}{n}\sum\left(\hat{\mathbf{u}}_{gt} - \mathbf{u}_{gt}\right)^2 \tag{9}$$

where $\hat{\mathbf{u}}_{gt}$ denotes the enhanced output. n is the number of pixels. The perceptual similarity assesses a solution concerning perceptually relevant characteristics. Here, the perceptual similarity is defined as the euclidean distance between the feature representations of enhanced images and clean instances. It can be formulated as follows:

$$\ell_{\text{PS}} = \frac{1}{m}\sum\left(\varphi_{i,j}\left(\hat{\mathbf{u}}_{gt}\right) - \varphi_{i,j}\left(\mathbf{u}_{gt}\right)\right)^2 \tag{10}$$

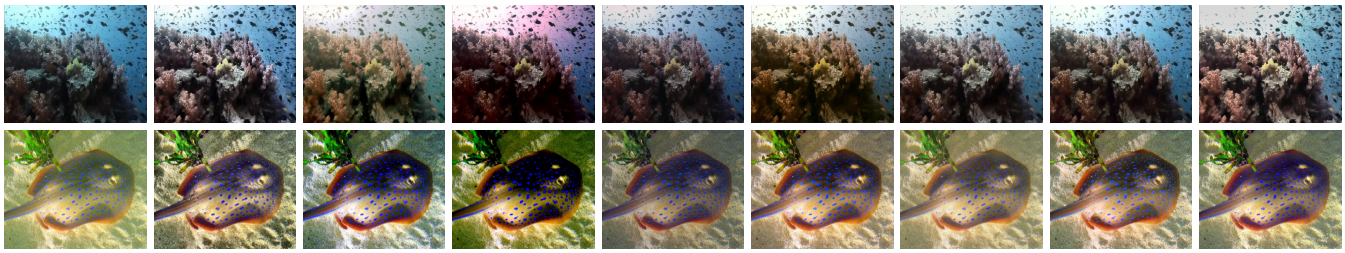
where $\varphi_{i,j}$ indicates the feature map obtained by the j-th convolution (after activation) before the i-th max-pooling layer within the pr-trained VGG16 network [19]. m is the number of pixels of all feature map extracted. The overall loss function consists of two components and is minimized during the network training. It is expressed as:

$$\ell_{\text{all}} = \ell_{\text{MSE}} + \lambda\ell_{\text{PS}} \tag{11}$$

where $\lambda = 0.1$ denotes the weight.

3. EXPERIMENT

We employ the real-world UIE dataset (UIEBD) [10] to train and test our model. UIEBD contains 890 underwater images and corresponding labels. Note that the reference image in UIEBD is obtained by subjective selections among 12 UIE results. We use the first 700 images for training and the rest for testing. We adopt the PyTorch framework to train our network using Adam solver with an initial learning rate of 1e-4. The mini-batch size is set as 1 empirically. The patch size is 128×128. The scale of the U-Net is set as 4 (i.e., we down-sample the features 3 times). We compare SCNet with six state-of-the-art UIE methods including one model-free approaches (Fusion [3]), two prior-based approaches (Histogram-Prior [6] and IBLA [20]), and three data-driven approaches (Water-Net [10], GLCHE [11], and Ucolor [12]). We use three objective quality assessment metrics (i.e., SSIM, PSNR, and LPIPS [21]) as the performance criteria. A better UIE approach should have higher SSIM and PSNR scores, and a lower LPIPS score. Besides, we test the model performance on another real-world dataset (RUIE) [1] to further



(a) Raw (b) Fusion (c) Histogram-Prior (d) IBLA (e) Water-Net (f) GLCHE (g) Ucolor (h) SCNet (i) Reference

Fig. 3. Visual comparisons on the UIEBD dataset.

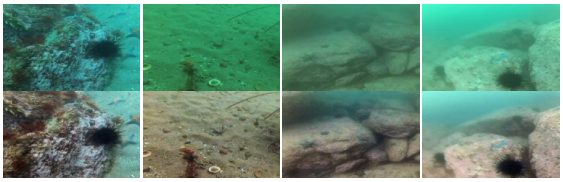


Fig. 4. Visual results on the RUIE dataset.

Table 1. Performance comparisons on the UIEBD dataset.

Method	SSIM ↑	PSNR ↑	LPIPS ↓
Fusion	0.8222	21.1849	0.2083
Histogram-Prior	0.7620	18.5148	0.3360
IBLA	0.5733	14.3856	0.4299
Water-Net	0.8303	19.3134	0.2016
GLCHE	0.8487	21.0270	0.1993
Ucolor	0.8395	21.6463	0.1970
SCNet	**0.8625**	**22.0816**	**0.1936**

Table 2. Ablation study on the UIEBD dataset.

Method	SSIM ↑	PSNR ↑	LPIPS ↓
U-Net	0.8350	19.5114	0.2640
SCNet w/o SN	0.8436	20.9934	0.2155
SCNet w/o CN	0.8579	21.3433	0.2070
SCNet (FULL)	**0.8625**	**22.0816**	**0.1936**

demonstrate the model generalization performance. Note that RUIE does not contain reference images. Therefore, only qualitative results are presented on this dataset.

3.1. Performance Comparisons

Quantitative results of different UIE algorithms on the UIEBD dataset are presented in Table 1. As we can observe, SCNet achieves the best performance in terms of three full-reference image quality evaluation metrics. Prior-based methods obtain low SSIM and PSNR scores. The reason may line in that prior-based methods rely on handcrafted imaging models and prior features. Model-free methods and data-driven methods achieve higher performance compared with prior-based methods, but all of them are inferior to our method since they ignore the impact of diverse water types. We also provide subjective comparisons in Fig. 3. From the visual results, we observe that SCNet can handle the diversity of water types, and can consistently generate natural and vivid results on testing images. Fig. 4 reports the visual results on the RUIE dataset. As can be observed, SCNet enables generating enhanced images with natural colors and contrasts. This demonstrates that the proposed method has good generalization performance for real-world applications.

3.2. Ablation Study

We conduct ablation studies to verify the effectiveness of the proposed spatial-wise and channel-wise normalization schemes. Table 2 presents the test results using four different settings. Due to the limited space, we use "SN" and "CN" to represent "Spatial-wise Normalization" and "Channel-wise Normalization," respectively. We can observe that directly using U-Net cannot obtain satisfactory results because it does not take the special distortions of the underwater environment into account. Normalizing representations on either spatial or channel dimensions can significantly improve enhancement performance. As expected, the best results are obtained by simultaneously normalizing features on both spatial and channel dimensions. This is because the diversity of water types exists in both spatial and channel dimensions.

4. CONCLUSION

We propose a novel data-driven method for underwater image enhancement. Different from most existing approaches that focus on designing complex network architectures, we propose to combine a simple U-Net with spatial-wise and channel-wise normalization to deal with the diversity of water types. By normalizing activations across both spatial and channel dimensions, appearance irrelevant representations can be effectively learned. As a result, the network can easily reconstruct the clean signal from those latent representations. Experimental results show that SCNet achieves competitive performance and has better generalization ability.

5. REFERENCES

[1] Risheng Liu, Xin Fan, Ming Zhu, Minjun Hou, and Zhongxuan Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4861–4875, 2020.

[2] Xueyang Fu, Peixian Zhuang, Yue Huang, Yinghao Liao, Xiao-Ping Zhang, and Xinghao Ding, "A retinex-based enhancing approach for single underwater image," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 4572–4576.

[3] Cosmin Ancuti, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert, "Enhancing underwater images and videos by fusion," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 81–88.

[4] John Y. Chiang and Ying-Ching Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1756–1769, 2012.

[5] Paulo L.J. Drews, Erickson R. Nascimento, Silvia S.C. Botelho, and Mario Fernando Montenegro Campos, "Underwater depth estimation and image restoration based on single images," *IEEE Computer Graphics and Applications*, vol. 36, no. 2, pp. 24–35, 2016.

[6] Chong-Yi Li, Ji-Chang Guo, Run-Min Cong, Yan-Wei Pang, and Bo Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5664–5677, 2016.

[7] Dana Berman, Tali Treibitz, and Shai Avidan, "Diving into haze-lines: Color restoration of underwater images," in *Proc. British Machine Vision Conference (BMVC)*, 2017, vol. 1.

[8] Derya Akkaynak and Tali Treibitz, "Sea-thru: A method for removing water from underwater images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1682–1691.

[9] Jie Li, Katherine A. Skinner, Ryan M. Eustice, and Matthew Johnson-Roberson, "Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 387–394, 2018.

[10] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2020.

[11] Xueyang Fu and Xiangyong Cao, "Underwater image enhancement with global-local networks and compressed-histogram equalization," *Signal Processing: Image Communication*, p. 115892, 2020.

[12] Chongyi Li, Saeed Anwar, Junhui Hou, Runmin Cong, Chunle Guo, and Wenqi Ren, "Underwater image enhancement via medium transmission-guided multi-color space embedding," *IEEE Transactions on Image Processing*, vol. 30, pp. 4985–5000, 2021.

[13] Saeed Anwar, Chongyi Li, and Fatih Porikli, "Deep underwater image enhancement," *arXiv preprint arXiv:1807.03528*, 2018.

[14] Pritish M Uplavikar, Zhenyu Wu, and Zhangyang Wang, "All-in-one underwater image enhancement using domain-adversarial learning.," in *CVPR Workshops*, 2019, pp. 1–8.

[15] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[16] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo, "Switchable whitening for deep representation learning," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1863–1871.

[17] Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie, "Positional normalization," pp. 1622–1634, 2019.

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] Yan-Tsung Peng and Pamela C. Cosman, "Underwater image restoration based on image blurriness and light absorption," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1579–1594, 2017.

[21] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.