JE^{2}NET: JOINT EXPLOITATION AND EXPLORATION IN REINFORCEMENT LEARNING BASED IMAGE RESTORATION

Xiaoyu Zhang[1,3], *Wei Gao*[1,3*], *Hui Yuan*[2] *and Ge Li*[1,3]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2]School of Control Science and Engineering, Shandong University, Jinan, China
[3]Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Previous reinforcement learning (RL) based image restoration studies typically train RL agents to search for recovery tools from a constructed toolset and iteratively recover images. However, we argue that these agents rely on pre-trained RL models with fixed-length paths for restoration, which performs poorly in the case of unknown distortions. To address these issues, we propose a joint exploitation and exploration reinforcement learning network (JE^{2}Net). Specifically, we propose a new deep classification network for image feature extraction and tool selection, which serves as a model prior. Second, we design a stochastic strategy to randomly select tools and a dynamic termination strategy to adaptively stop the recovery process. In this way, the model prior and exploration mechanism can be jointly used to expand the search space and obtain more quality gain. Experimental results show that our proposed method is more flexible compared to other state-of-the-art methods and achieves significant quality improvements in the presence of unknown distortions.

Index Terms— Image Restoration, Reinforcement Learning, Deep Learning

1. INTRODUCTION

As a fundamental problem in low-level computer vision, image restoration [1] aims to remove artifacts for better visual experience, or better performances of subsequent vision tasks. Images in the real world are subject to a variety of distortions due to diversified imaging devices, complex external environments, and unknown degradation processes.

According to the characteristics of degradation processes, traditional methods usually firstly establish degradation models, and design different filtering methods [2, 3]. Due to robustness and adaptive learning capabilities, deep convolutional neural networks (CNNs) have been investigated for artifacts removal, which have ameliorated performances significantly. The CNN-based networks [4, 5, 6, 7, 8, 9, 10, 11, 12, 13] predict the input degraded images by a well-designed network and output the recovered high-quality images. Nevertheless, these methods perform image recovery based on trained models, thus relying on the trained dataset. If the order of distortion or the degree of distortion is changed, the CNN-based methods need to be completely retrained, resulting in poor generalization under unknown distortions. Recently, deep reinforcement learning (RL) based methods [14, 15, 16] are used for image restoration. These approaches typically train a policy network as an agent, and the RL agent iteratively selects tools from the toolset and applies them to the images to achieve recovery. Such methods exemplify the potentials of using multiple recovery tools jointly. However, they are less flexible in the face of unknown combinatorial deformations as they rely on the results of the network output and use a fixed number of enhancement steps, leading to a bottleneck of recovery performance.

From the above analysis, in order to efficiently handle various unknown distortions and fully exploit the potential of multi-step enhancement, in this paper, we propose a new RL-based image restoration architecture, namely joint exploitation and exploration reinforcement learning network (JE^{2}Net). **Firstly**, to handle various types of distortions, we propose a new toolset containing ten CNN-based restoration tools to handle different distortions. **Secondly**, we design a policy network based on ResNet to perform tool selection. It is updated by deep Q-learning to select those tools that maximize the long-term total return. **Thirdly**, to deal with unknown distortions, we introduce the epsilon greedy algorithm to randomly select tools with a certain probability. It enable us to select tools in a larger search space rather than being confined to the selection of the policy network. **Finally**, we design a dynamic strategy to terminate the iterations so as to fully explore in the new search space. In case of more severe distortions, the framework adaptively takes longer steps to recover the image adequately. In this way, the proposed

*Corresponding author: gaowei262@pku.edu.cn

This work was supported by Natural Science Foundation of China (61801303, 62031013), Guangdong Basic and Applied Basic Research Foundation (2019A1515012031), Shenzhen Fundamental Research Program (GXWD20201231165807007-20200806163656003), Shenzhen Science and Technology Plan Basic Research Project (JCYJ20190808161805519), and CCF-Tencent Open Fund (RAGR20200114).

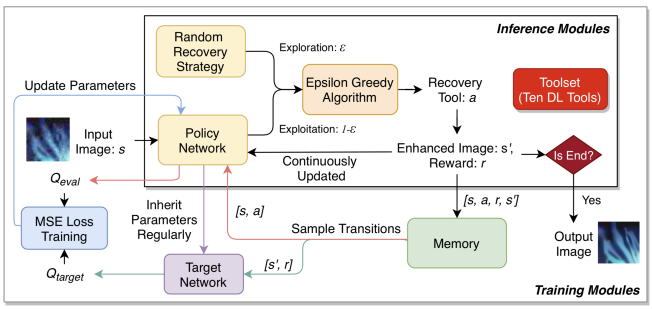


Fig. 1. Illustration of the proposed deep reinforcement learning based image restoration framework, including the details of training and inference process.

exploration mechanism incorporates a stochastic scheme with a dynamic termination strategy, which is more flexible when encountering different unknown distortion scenarios and can achieve better quality gain.

Our main contributions are summarized as follows. (1) We propose a joint exploitation and exploration framework (JE^{2}Net) for online image restoration, which can effectively cope with unknown distortion processes. (2) We introduce epsilon greedy algorithm and combine it with an adaptive stopping criterion that can dynamically find better restoration tools. (3) Experimental results show that the proposed framework achieves significant gain in a variety of distortion scenarios and outperforms other state-of-the-art methods.

2. PROPOSED METHOD

2.1. Recovery Toolset

As shown in Table 1, we select a series of efficient networks for restoration. For denoising, we use DnCNN [4] and FFDNet [5] networks, where the variance of the noise distribution of FFDNet is 15 and 25, respectively. For super resolution, we reproduce IMDN [6], ESRGAN [7] (both MSE-loss and GAN-Loss) and VDSR [8] networks. Moreover, SRN [10] is used for reduce motion blurring. As for JPEG compression artifacts removal, we introduce DMCNN [9] training with quality factors (QFs) of 10 and 20, respectively. Therefore, the proposed toolset consisting of the above ten tools has the potential to handle different types of distortions.

2.2. Deep Reinforcement Learning Framework

The proposed framework is based on deep Q-learning network (DQN) [17], which is illustrated as Fig. 1. RL agents usually select actions for each state according to a policy and receive the corresponding rewards. Afterwards, the policy is updated accordingly. In our work, the policy network is implemented by Resnet [18]. States are represented by images, actions are tools described in Table 1, and rewards are

Table 1. Candidate recovery tools in the proposed toolset.

IDs	Types	Implementations
0	Denoise	DnCNN
1		FFDNet15
2		FFDNet25
3	Super-Resolution	IMDN
4		ESRGAN$_{MSE}$
5		ESRGAN$_{GAN}$
6		VDSR
7	Deblur	SRN
8	DeJPEG	DMCNN$_{QF10}$
9		DMCNN$_{QF20}$

quantified by peak signal-to-noise ratio (PSNR). Moreover, we implement a target network with Resnet to make the training more stable. The policy network extracts image features and output Q values for each recovery tool, which indicates the estimated long-term value of the tool. Then, the epsilon greedy algorithm [19] is adopted, i.e., the tool with the largest Q value output is selected with probability 1-ε, or the tool is randomly selected. Afterwards, we update the distorted image s with the selected tool a to get an enhanced image s' and the reward r_s^a. The two images, action and reward are stored as a transition state $[s, a, r_s^a, s']$. After storing a certain number of transitions, we sample some transitions randomly from memory to update the policy network by minimizing:

$$L = \frac{1}{2N} \sum_{i \in N} \| r_s^a + \gamma \max Q^*(s', :) - Q(s, a) \|_2^2, \qquad (1)$$

where N refers to the number of samples, r_s^a is the PSNR gain obtained after applying tool a to the image s. γ is an attenuation factor. $Q(s, a)$ represents the estimated value of tool a for image s from policy network. $Q^*(s', :)$ represents all the estimated values for image s' from target network. Additionally, the target network inherits the parameters of the policy network with regular intervals. After such training, we can have an policy model for further action selection.

2.3. Proposed Exploration Mechanism

For online inference, we consider both the policy model and exploration scheme. Tools with the largest probabilities are regarded as the model selections, named exploitation. While exploration is reflected by larger probability ε and longer enhancement trails. To further demonstrate the important role of exploration in quality enhancement, we degenerate seven images from ImageNet [20] by sequentially applying bicubic downsampling with scale of 4, JPEG compression with a quality factor of 10, and additive Gaussian noise which distributed at 15. Two different schemes are experimented in Table 2, including "Try-All-NEpsilon" which attempts all kinds of 1-ε from 0 to 1.0 with intervals of 0.1, and "Only-Use-Net" which only uses the policy network for action selection. It can

Table 2. Performance comparisons for seven distorted images with enhancement trails of 15. Recovery tool IDs are sequentially given.

Images	Try-All-NEpsilon		Only-Use-Net	
	gain	Paths	gain	Paths
1	1.9270	115307	1.8943	14
2	2.0174	146744381004000	1.9342	164644
3	3.2177	116662100300000	2.3752	166262
4	2.1146	1461610	2.0289	146
5	0.9110	71274473	0.8586	36
6	1.1391	71234	1.0181	174
7	1.9987	714627461400007	1.4583	1646

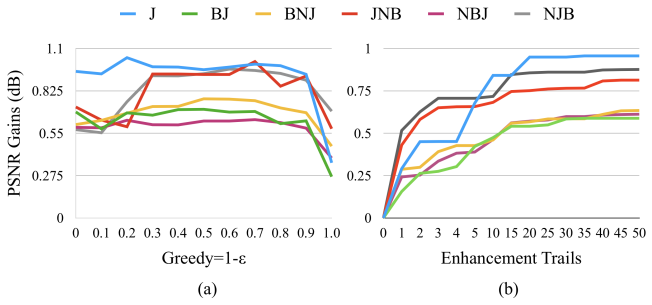


Fig. 2. Trends of PSNR gain (dB) for different types of distorted images. (a) Different greedy parameters with trails=50. (b) Different enhancement trails with greedy=0.7.

be seen that the use of randomness allows searching over a larger space and fully exploiting the potential for quality improvement, thus contributing to better overall performance.

To obtain the optimal 1-ε value and the length of the enhancement trajectory, we experiment on various distortion scenarios, including single, double and triple distortion. Specifically, we enumerate the distribution of PSNR gain under different 1-ε parameters and different enhancement trajectories. As shown in Figure 2, the PSNR gain varies quadratically with 1-ε, and reaches a peak at around 0.7 on average. For the length of enhancement trails, longer trails are beneficial to obtain better image quality, and the PSNR gain reaches convergence at about 50 steps. However, since images with different levels of distortion usually require different enhancement path lengths, we set a dynamic stopping criterion, i.e., stopping recovery when there is no quality gain in three consecutive steps.

3. EXPERIMENTS

3.1. Experimental Setup

We perform experiments on 800 high-resolution images from DIV2K [21] dataset for policy network training. The training images are randomly degenerated by applying bicubic downsampling (B) with scale of 4, JPEG compression (J) with a quality factor of 10 and gaussian noise (N) uniformly distributed in [0, 15]. For inference, Set5, Set14 [22] and the DIV2K test set [21] datasets with a total of 119 high-resolution images are used. Different degradation factors (i.e., B,J,N) are combined using one, two or three of different distortion types, and therefore the generalization capability of different methods can be tested on diversified cases. In addition, we conduct experiments on 3000 images from TID2013 [23], containing 24 distortions as well as 5 distortion levels. The wild dataset CLIVE [24] is experimented for unknown distortions, which was taken in a real wild environment.

It is pivotal to set greedy parameter 1-ε properly for both training and exploration-exploitation tradeoff. As for training, a small greedy will bring benefits to random exploration, but it will also make convergence difficult. Hence, we set 1-ε=0.5 and increase it by 0.1 every 100 epochs until 0.9 to balance exploration with convergence. For inference, 1-ε is set to 0.7 as mentioned in Section 2.3. Besides, we set enhancement trails as 50, which is sufficient to fully explore the recovery path, and the enhancement trajectory is terminated by the stopping criterion. Moreover, the memory size and sampling numbers are set to 100 and 30, and γ is set to 0.9. The target network inherit parameters every 30 times.

3.2. Performance Comparisons

The quantitative results are presented in Table 3, it can be seen that our proposed method outperform all other state-of-the-art methods. Compared to FFDNet [5], VDSR [8] and DMCNN [9] using a single convolutional network, our adaptive use of multiple tools can increase PSNR and SSIM gain by 0.372dB and 0.0385, respectively. Compared with other RL-based network (i.e., RL-Restore [14] and PixelRL [15]), the proposed framework with exploration-exploitation tradeoff and longer enhancement trails could benefit from the further potential gain, and the achieved PSNR and SSIM improvements are 0.995dB and 0.0703, respectively.

3.3. Visualization Comparison.

We visualize some trails in enhancement process as Fig. 3. From the superposition and alternating use of multiple tools, image quality can be gradually improved as enhancement trials increase. This also validates that the proposed longer enhancement trails can be beneficial to quality gain. Some visualization results on DIV2K and CLIVE are given in (a) and (b) of Fig. 4, respectively. It is observed that the proposed method performs better in blocking artifacts and noise removal, and reproduces fine texture details.

3.4. Ablation Study

To validate the effectiveness of the joint use of the policy model and randomness, we compare the results with and without either of them on triple distortion datasets of DIV2K in

Table 3. Comparisons of PSNR (dB) and SSIM gain on {Set5, Set14, DIV2K} with different distortion types and TID2013.

Distortion Types	FFDNet		VDSR		DMCNN		PixelRL		RL-Restore		JE^{2}Net	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
B	-0.105	-0.0204	0.574	0.0681	-0.045	-0.0099	0.105	0.0185	0.307	0.0407	0.821	0.0736
J	0.454	0.0206	-1.639	-0.0485	0.638	0.0215	0.131	0.0069	0.824	0.0275	0.743	0.0207
N	4.064	0.1011	-2.472	-0.0668	-3.964	-0.0725	2.541	0.0499	2.625	0.0619	3.294	0.0872
Average	1.471	0.0337	-1.179	-0.0157	-1.124	-0.0203	0.925	0.0251	1.252	0.0434	**1.620**	**0.0605**
BJ	0.139	0.0085	-0.108	-0.0059	0.083	0.0047	0.080	0.0127	0.449	0.0454	0.482	0.0458
JB	-0.009	-0.0084	0.448	0.0615	-0.017	-0.0050	0.055	0.0184	0.394	0.0443	0.576	0.0538
BN	0.952	0.0940	-0.470	-0.0264	-0.121	-0.0101	0.953	0.0701	1.217	0.1214	1.456	0.1563
NB	-0.083	-0.0167	0.540	0.0675	-0.039	-0.0077	0.106	0.0226	0.324	0.0440	0.764	0.0696
JN	1.479	0.1043	-1.342	-0.0411	0.179	-0.0006	1.318	0.0655	1.887	0.1022	1.613	0.0950
NJ	0.486	0.0239	-1.708	-0.0531	0.695	0.0270	0.131	0.0076	0.892	0.0324	0.807	0.0243
Average	0.494	0.0343	-0.440	0.0004	0.130	0.0014	0.441	0.0328	0.861	0.0650	**0.950**	**0.0741**
BJN	0.957	0.0932	-0.430	-0.0187	-0.050	-0.0108	0.900	0.0614	1.232	0.1190	1.330	0.1366
BNJ	0.170	0.0123	-0.159	-0.0143	0.111	0.0082	0.073	0.0124	0.463	0.0444	0.526	0.0477
JBN	0.905	0.0901	-0.433	-0.0005	-0.104	-0.0095	0.857	0.0617	1.184	0.1244	1.280	0.1378
JNB	0.001	-0.0068	0.430	0.0426	-0.016	-0.0044	0.059	0.0190	0.411	0.0471	0.582	0.0560
NBJ	0.143	0.0089	-0.112	0.0129	0.087	0.0051	0.080	0.0133	0.453	0.0471	0.520	0.0491
NJB	-0.001	-0.0073	0.437	0.0438	-0.012	-0.0039	0.053	0.0193	0.400	0.0450	0.573	0.0552
Average	0.362	0.0317	-0.044	0.0110	0.003	-0.0025	0.337	0.0312	0.691	0.0712	**0.802**	**0.0804**
TID2013	0.418	0.0117	0.002	0.0001	0.0003	-0.0001	-1.671	0.0000	0.398	0.0129	**0.550**	**0.0160**
Total Average	0.623	0.0318	-0.403	0.0013	-0.161	-0.0043	0.361	0.0287	0.841	0.0600	**0.995**	**0.0703**

Table 4. PSNR gain (dB) for ablation study on various distortions.

Schemes	BJN	BNJ	JBN	JNB	NBJ	NJB
w/o PolicyNet	1.209	0.466	1.309	0.579	0.389	0.693
w/o Random	1.708	0.608	1.856	0.722	0.588	0.570
Path Lengths=3	1.390	0.349	1.507	0.582	0.333	0.635
Path Lengths=4	1.458	0.419	1.570	0.669	0.356	0.720
Path Lengths=5	1.557	0.447	1.605	0.729	0.387	0.797
Proposed	**1.758**	**0.761**	**1.940**	**1.023**	**0.638**	**0.957**

Table 4. As can be seen from the table, the PSNR gain on multiple distortions decreases by 0.171dB when only using the policy network, while it decreases by 0.406dB when selecting the tool completely at random, indicating that they both contribute to the quality improvements. To illustrate the superiority of longer recovery steps, we conduct experiments with limited path lengths of 3, 4, and 5. As can be observed in Table 4, the PSNR gain gradually increases to 0.799dB, 0.865dB and 0.920dB as the path length increases, and finally reaches 1.180dB. Therefore, the proposed sufficiently enhancement trails and the dynamic stop criteria help to perform adequate exploration and bring the most available gain.

4. CONCLUSION

In this paper, we propose a novel deep reinforcement learning-based framework (JE^{2}Net) to efficiently remove single and multiple artifacts. Firstly, a deep policy network and an exploration mechanism are jointly used to explore the best recovery path. Specifically, a deep policy network with a diversified toolset is trained in advance to select tools. Then, a stochastic strategy is subsequently applied with a certain probability to decide the final tool to be taken. Finally, the recovery process is dynamically stopped according to the termination criteria. The proposed exploration mechanism can effectively expand the search space for more potential gain. Extensive experimental results demonstrate the superiority and flexibility of our proposed method over the other state-of-the-art methods in the face of unknown distortions.

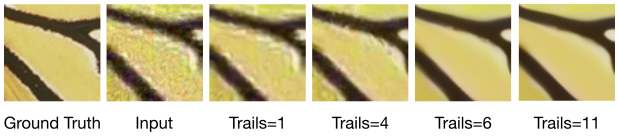
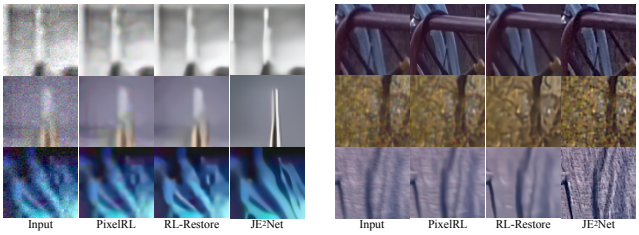


Fig. 3. Incremental enhancement process.



(a) Comparisons on DIV2K. (b) Comparisons on CLIVE.

Fig. 4. Comparisons of visualized image quality for different approaches on various datasets.

5. REFERENCES

[1] Lanqing Guo, Zhiyuan Zha, Saiprasad Ravishankar, and Bihan Wen, "Self-convolution: A highly-efficient operator for non-local image restoration," in *ICASSP*, 2021, pp. 1860–1864.

[2] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel, "Non-local means denoising," *Image Processing On Line*, vol. 1, pp. 208–212, 2011.

[3] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, 2007.

[4] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017.

[5] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Ffdnet: Toward a fast and flexible solution for cnn-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, 2018.

[6] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang, "Lightweight image super-resolution with information multi-distillation network," in *ACMMM*, 2019, pp. 2024–2032.

[7] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *ECCV*, 2018, pp. 0–0.

[8] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016, pp. 1646–1654.

[9] Xiaoshuai Zhang, Wenhan Yang, Yueyu Hu, and Jiaying Liu, "Dmcnn: Dual-domain multi-scale convolutional neural network for compression artifacts removal," in *ICIP*, 2018, pp. 390–394.

[10] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia, "Scale-recurrent network for deep image deblurring," in *CVPR*, 2018, pp. 8174–8182.

[11] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," in *ICCV*, 2019, pp. 8878–8887.

[12] Masanori Suganuma, Xing Liu, and Takayuki Okatani, "Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions," in *CVPR*, 2019, pp. 9039–9048.

[13] Guocheng Qian, Jinjin Gu, Jimmy S Ren, Chao Dong, Furong Zhao, and Juan Lin, "Trinity of pixel enhancement: a joint solution for demosaicking, denoising and super-resolution," *arXiv preprint arXiv:1905.02538*, 2019.

[14] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy, "Crafting a toolchain for image restoration by deep reinforcement learning," in *CVPR*, 2018, pp. 2443–2452.

[15] Ryosuke Furuta, Naoto Inoue, and Toshihiko Yamasaki, "Pixelrl: Fully convolutional network with reinforcement learning for image processing," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1704–1719, 2020.

[16] Ke Yu, Xintao Wang, Chao Dong, Xiaoou Tang, and Chen Change Loy, "Path-restore: Learning network path selection for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[17] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[19] James D. McCaffrey, "The epsilon-greedy algorithm," 2017, https://jamesmccaffrey.wordpress.com/2017/11/30/the-epsilon-greedy-algorithm/.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[21] Eirikur Agustsson and Radu Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *CVPRW*, 2017, pp. 126–135.

[22] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie line Alberi Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *BMVC*, 2012, pp. 135.1–135.10.

[23] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al., "Image database tid2013: Peculiarities, results and perspectives," *Signal processing: Image communication*, vol. 30, pp. 57–77, 2015.

[24] Deepti Ghadiyaram and Alan C Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, 2015.